BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

The Quality of Google Translate and ChatGPT English to Arabic Translation: The Case of Scientific Text Translation

***Elham Alzain*[1*] , *Khalil A. Nagi*[2] , *Faiz Algobaei*[3]**

[1]*King Faisal University, Alahsa, Saudi Arabia*

[2]*University of Saba Region, Marib, Yemen*

[3]*Northern Border University, Rafha, Saudi Arabia*

ABSTRACT

The aim of the study is to investigate the quality of neural machine translation (NMT) and that of large language models (LLMs). The research team uses Google Translate and ChatGPT in the translation of various selected scientific texts. They provide an evaluation of the translation outputs. Both an error analysis and human evaluation are provided by professional annotators. The error analysis is provided based on the typology of errors introduced by Multidimensional Quality Metrics (MQM). A professional evaluation is also provided using a 7-point Likert scale. The professional annotators provide an evaluation on the document level. Both the evaluation and the number of errors show that Google Translate outperforms ChatGPT. However, the results indicate that both systems still require a lot of training. It is also suggested that annotated corpora need to be constructed. The study provides invaluable insights on the strength and weakness of the systems under study which will be beneficial for translators, researchers and developers of machine translations.

Keywords: Google Translate; ChatGPT; Translation quality; Error analysis; English-Arabic

*CORRESPONDING AUTHOR:

Elham Alzain, King Faisal University, Alahsa, Saudi Arabia; Email: elhamalzain@gmail.com

ARTICLE INFO

Received: 26 June 2024 | Revised: 22 July 2024 | Accepted: 29 July 2024 | Published Online: 27 August 2024
DOI: https://doi.org/10.30564/fls.v6i3.6799

CITATION

Alzain E., Nagi K.A., Algobaei, F., 2024. The Quality of Google Translate and ChatGPT English to Arabic Translation: The Case of Scientific Text Translation. Forum for Linguistic Studies. 6(4): 837-849. DOI: https://doi.org/10.30564/fls.v6i3.6799

1. Introduction

Machine translation (MT) is a field that exhibits an astonishing and rapid development. With such development, there is a lot of debate regarding the quality of translation outputs. Some researchers believe MT has reached a point where it becomes in par with professional human translation (Hassan et al., 2018; Barrault et al 2019, among others). Others, however, argue that MT still falls short when compared to professional translation (Läubli et al, 2018; Toral et al 2018; Freitag et al, 2021, among others).

However, regardless of these ongoing debates, it is evident that machine translation is progressing and high-quality translation outputs are produced. Nonetheless, it is still clear that there is still a gap between the quality of machine translation and that of professional human translation. Studies conducting error analyses have identified a considerable number of errors that fall under different categories (Popović, 2021; Kocmi et al., 2022; Nagi, 2023; Almekhlafi & Nagi, 2024, among others). It should be mentioned that the variation and number of errors are more noticeable when a language with poor morphology is translated into a language with richer morphology, as in the case of translating English into Arabic (Nagi, 2023; Almekhlafi & Nagi, 2024).

With the advent of large language models (LLMs), a new wave of research has started for the purpose of investigating the translation performance of LLMs such as ChatGPT and Brad. GPT models perform well on high-resource languages where they compete with commercial translation systems like Google Translate, but they fall behind when dealing with low-resource languages (Hendy et al, 2023; Jiao et al 2023).

Hendy et al. (2023) have stated that the performance of GPT models is under-investigated compared to commercial systems. It is also indicated that the improvement of automatic translation requires more fine-grained analyses in regard to translation quality (Popović, 2021; Kocmi, 2022). It is, therefore, an interesting topic to investigate the translation quality of LLMs and compare it to that of NMT.

This study aims to investigate the translation quality of Google Translate and ChatGPT when translating selected English scientific texts to Arabic. It also presents an evaluation of the quality of the translation outputs produced by both systems. The study also aims to introduce a classification of errors that occur when translating English scientific texts to Arabic using Google Translate and ChatGPT.

The evaluation presented in this study is based on seven-point-Likert-like scale by professional annotators. Errors are classified according to the Multidimensional Quality Metrics (MQM) typology. The analysis provides the nature of the errors that occur when translating English scientific texts to Arabic with explanations and examples.

The study, therefore, contributes greatly to the efforts of systematically evaluating the translation quality of both NMT systems and LLMs. To the best of the researchers' knowledges, this is the first study that compares the translation outputs of Google Translate and ChatGPT that work on MT of scientific texts from English to Arabic. Even, studies on other language pairs have provided general comparisons and not specified to scientific texts. In addition, previous studies have shown that GPT models fall behind when a low-source language is involved. Therefore, it is important to provide a detailed analysis and evaluation on how these models perform when providing Arabic translation. The significance of the study also comes from the fact that Arabic is a morphologically rich language which forms a challenge for MT systems in addition to the nature of the texts used in the study.

It should be mentioned here that this study is limited by its focus on English-Arabic translation of scientific texts using Google Translate and ChatGPT and the results may not be applicable to other language pairs or other text genres. The size of the investigated texts seems to be limited. However, with the in-depth error analysis provided in the study, the researchers believe that the potential errors to appear in the translation outputs of the systems under investigation are covered. That is evident from the long list of the annotated error types.

2. Literature review

2.1 Neural machine translation and predecessors

Machine translation (MT) has received significant interest in the literature of language research and the development it has achieved recently is remarkable. The introduction of Neural Machine Translation (NMT) systems was a major breakthrough in the MT field and it has prompted enormous research to evaluate their performance. Researchers evaluate NMT outputs and compare the translation quality of NMT

systems to that of previous systems to identify the strengths and weaknesses of such systems. Such investigations are performed to contribute to the development of MT systems.

Various research works comparing the performance of Neural Machine Translation (NMT) and Phrase-Based Machine Translation (PBMT) systems have concluded that NMT outperforms PBMT in different areas. Bentivogli et al. (2016) have analyzed English to German translation outputs and found that NMT reduces the post-editing efforts and significantly improves the inflection and word order of the output. Toral and Sanchez-Cartagena (2017) have also conducted analyses on the performance of NMT and PBMT and concluded that NMT is better in terms of inter-system variability, producing fluent outputs, and re-ordering. Similarly, Klubicka et al. (2017) examined English to Czech translation output produced by both systems reaching the conclusion that NMT surpasses PBMT when it comes to producing fluent and grammatical language and handling agreement. However, NMT's performance degrades more quickly in the case of longer sentences (Bentivogli et al., 2016; Koehn and Knowles, 2017).

According to recent research, NMT systems outperform other Statistical Machine Translation (SMT) systems as well. Sennrich and Zhang (2019) have concluded that NMT systems are better when low-resource languages and general domains are involved. Ahmadnia and Dorr (2020) have also proposed that NMT surpasses SMT in the case of low-resource domains with specific data. In addition, Saunders (2022) has noted that NMT systems benefit significantly from domain adaptation in achieving better performance with less training data.

Regardless, the quality of NMT outputs remains a controversial issue. From one standpoint, some studies have claimed that MT systems have advanced to a level where an MT output is almost on par with human translation. Isabelle et al. (2017) have stated that NMT is nearly on par with human translation especially when it comes to close language pairs such as English and Spanish or English and French. Levin et al. (2017) have also concluded that NMT achieves fluency that is comparable to human translation in the cases of English into German translation and English into French translation. Moreover, Hassan et al. (2018) and Popel et al. (2020) have also concluded that, in some specific cases, machine translation has matched or even surpassed professional human translation.

Nevertheless, there is significant evidence suggesting that the gap between professional human translation and machine translation is still substantial and that machine-human parity is not achieved yet (Toral et al., 2018; and Freitag et al., 2021). Other analyses have also highlighted the nature and types of various errors that appears in machine translation outputs with an emphasis on the need for more detailed error studies (Daems et al., 2014; Popović, 2021; Kocmi et al., 2022; Rivera-Trigueros, 2022; Nagi, 2023; Almekhlafi and Nagi, 2024, among others). Such fine-grained analyses are crucial since they provide insights into the strengths and weaknesses of MT systems and help in the development of MT systems and in the facilitation of post-editing.

Now with the occurrence of LLMs, detailed analysis of error types and evaluation of translation quality are required. It could, therefore, be beneficial to compare the performance of the new models with the performance of the preceding NMT systems as it will be carried out in this study. Let us, therefore, provide an overview of the new model to be used in this study, ChatGPT.

2.2 ChatGPT

ChatGPT is the most well-known artificial intelligence (AI) application nowadays. It gained its fame even before its advent, as people anticipated its usage in daily life when some companies, such as BBC, CNN, and People's Daily, announced the upcoming AI revolution. According to Siu (2023), ChatGPT's rapid popularity is due to the idea that it can perform many tasks such as generating texts, answering questions, classifying texts, generating codes, and translating languages very well. The reason behind this is that ChatGPT employs many methods like Natural Language Processing (NLP), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Transformer and Reinforcement Learning from Human Feedback model (RLHF).

In general, ChatGPT's performance of these various tasks has been investigated. ChatGPT is presumed to produce outputs which show that it is indifferent to the truth (Hicks et al., 2024), or it shows inconsistent performance and bias (Buscemi & Proverbio, 2024). It shows issues in discourse parsing (Chan et al., 2024) or in detecting ambiguities (Ortega-Martín et al., 2023).

The performance of ChatGPT in translation is still de-

fective. According to Jiao et al. (2023) and Hendy et al. (2023), ChatGPT is as good as commercial translation systems like Google Translate when translating high-resource European languages, but it falls short when translating low-resource or distant languages. In addition, ChatGPT is not as good as commercial translation systems at translating biomedical abstracts or Reddit comments, but it is good at translating spoken language. However, the launch of the GPT-4 engine in March 2023 has significantly improved ChatGPT's translation performance, making it comparable to commercial translation systems even for distant languages (Siu, 2023).

In addition, Siu (2023) has stated that GPT-4 and ChatGPT are beneficial in translation since they could be used in different ways for enhancing translations' accuracy, clarity, and fluency by using specific prompts to detect errors, revise translations, provide word synonyms, etc. Those models could be requested to rewrite some texts in different ways or in a specific style. They have the ability to provide the accurate meaning of a multi-word expression with multiple meanings, based on the context of the text. According to the author, professional translators could take the advantage to work collaboratively well with those models to produce perfect translations.

Bubeck et al. (2023) have explained that there are some flaws in GPT-4, some of which could be attributed to its built in next-word prediction architecture. In addition, OpenAI (2023) has reported some other flaws such as unreliability, hallucination, limited knowledge, overconfidence in incorrect predictions, and insufficient self-checking. Jiao et al (2023) have also pointed out that GPT-3.5 model does not achieve the desired results in specific domains compared to its performance with spoken language translations. Hendy et al (2023) have pointed out that GPT have restricted potential in low-resource languages. Zhu et al. (2023) have also stated that multilingual translation capabilities of LLMs are improving with time and GPT-4 demonstrates exceptional performance capabilities. However, GPT-4 has not yet reached the desired level of performance with respect to low-resource languages.

Khoshafah (2023) has pointed out that ChatGPT cannot perform well in dealing with translations of specialized texts such as scientific, legal, medical texts, or literary works while it can do well in translating simple content. In her study, Khoshafah has focused on the accuracy of translating the aforementioned texts comparing it to human translation and presented examples of terminology errors. She also suggested using human translation to train ChatGPT. In that regard, Nagi et al. (2024) also concluded that ChatGPT faces a challenge and it produces outputs with high error frequency when translating complex sentences from English into Arabic.

It should be mentioned here that in the literature of MT, comparisons between the translation outputs of LLMs and NMT systems have been performed. Aghai (2024), for example, have conducted a study that focused on the quality of translating literary texts from Persian to English. The study presents an overall evaluation of different aspects of translation quality based on detected errors.

This, however, is the first study that compared the translation outputs of Google Translate and ChatGPT when translating scientific text from English to Arabic. The study evaluated the translation quality in addition to in-depth errors analysis and taxonomy. The study, therefore, provide better insight into the weakness and strength points of the systems under investigation when translating scientific texts.

2.3 Machine translation and scientific texts

Scientific text translation has been ignored in the literature since it requires more expertise and effort (Tehseen et al., 2018). It has been assumed that it is problematic due to the fact that there are various fields of science with their own terminology. Recent studies have also shown that machine translation outcomes exhibit various translation errors such as word order, absence of articles, word choice, and translating abbreviations (Ulitkin, 2021). Studies have also shown that machine translation of scientific texts needs very careful post editing (Zulfiqar et al, 2018; Escartín & Goulet, 2020). Escartín & Goulet (2020) have also presented a long list of post-edits that concern the fluency and adequacy of the translated texts.

Therefore, it should be noted here that scientific text translation is still in need of more investigation when it comes to neural machine translation. However, the matter is more serious when it comes to LLMs since the field of translating scientific texts using ChatGPT is still almost untouched. There are very few studies which have provided general evaluation of ChatGPT translation against other systems such as

the ones mentioned in Section 2.1 and Section 2.2. However, according to the researchers' knowledge, there is no study specified to using LLMs in translating scientific texts.

3. Methodology and results

3.1 Dataset and annotators

The research team uses Google Translate and ChatGPT to translate 33 English texts that contain 209 sentences. To ensure the variety of the texts, the researchers selected the texts from different resources and from three different science fields: medicine, biology and computer science. Though the number of sentences is limited, it fits the nature of the study. The researchers believe that they are enough to examine the issues that MT face when translating scientific texts. That is due to the fact that the texts were extracted from various resources and that they represent different scientific fields. In addition, the included sentences show variation in term of structural features. Besides, a human extensive analysis and taxonomy of error was performed and the texts were thoroughly examined.

Furthermore, to ensure the annotation quality, the analysis is carried out by two private professional translators who have a long experience in the field of translation and annotation. The first annotator has a ten-year experience whereas the second annotator has a seven-year experience in the respective field. Both annotators are native speakers of the target language (Arabic) and have near native fluency in the source language (English). The researchers perform a pilot study where three texts are provided with annotation guidelines to the annotators. The annotations are thoroughly reviewed by the research team and feedback is provided to the annotators in the case of an occurrence of a misunderstanding of the process or the guidelines. The research team also clarifies all doubts and answers all questions raised by the annotators.

3.2 Evaluation

Translation quality assessment (TQA) is a complex issue that has been debated by both academics and industry professionals. In academia, TQA is typically concerned with developing measures that can demonstrate a change in quality either by showing improvement in a translation compared to previous work or between different translation processes. However, in industry, the aim is to ensure that a specified level of quality is met (Castilho et al, 2018).

It should be noted here that evaluation of MT was mainly focused on sentence-level evaluation. However, recently the significance of document level evaluation has been brought into attention (Toral et al., 2018; Läubli et al., 2018; Läubli et al., 2020; Graham et al., 2020; Toral, 2020, among others). Accordingly, two main types of evaluation have been proposed: full document-level evaluation as proposed in Läubli et al. (2018) and Läubli et al. (2020) and segment-level evaluation as proposed in Graham et al. (2019) and Graham et al. (2020).

Läubli et al. (2018), however, proposed "pairwise ranking" of both fluency and adequacy where the two translations of each text pair were compared. The translation that adequately expressed the meaning of the source text was considered to have higher accuracy, and the translation with better language was considered to have higher fluency. Läubli et al. (2020), have also provided some recommendations that strengthen the efficiency of the evaluation. The framework introduced in Läubli et al. (2018) also provided concrete evidence against claims that MT are on par with human translation. Recommendations provided in Läubli et al. (2018) have adopted the large-scale evaluation campaign at WMT 2019 (Barrault et al., 2019). It is also indicated that Läubli et al.'s (2020) recommendations represent great progress in the evaluation field (Poibeau, 2022). It is also worth mentioning that Castilho (2020) has pointed out that this evaluation method is appropriate when comparing different translations and not when evaluating a single MT system.

Segment-level evaluation is also another method in the literature that was proposed to be efficient in this aspect. In segment-level evaluation, a direct assessment is provided for sampled segments (Graham et al., 2019; Graham et al., 2020). However, it is shown that segment-level evaluation shows tendency to minimize the difference between human translation and MT (Barrault et al., 2019; Läubli et al., 2020).

In this study, however, the purpose of the evaluation is not only to identify which system performs better, but to identify how good each system performs. The research team, therefore, performed a professional evaluation based on scalar quality metric (SQM) (Freitag et al., 2021). The study employed the SQM which uses a 0-6 Likert-like scale,

whose rankings are as follows.

- 6: Perfect Meaning and Grammar: The meaning of the translation is completely consistent with the source and the surrounding context (if applicable). The grammar is also correct.
- 4: Most Meaning Preserved and Few Grammar Mistakes: The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.
- 2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.
- 0: Nonsense/ No meaning preserved: Nearly all information is lost between the translation and source. Grammar is irrelevant.

The source texts were provided to the two professional annotators along with their correspondent translations. The annotators are presented with the SQM guidelines to provide an evaluation for each text. The annotators evaluated the translation outputs of the whole texts and not just selected segments. It should be noted that this method of evaluation is used in both WMT 2022 and WMT 2023 General Machine Translation Task (Kocmi et al, 2022, Kocmi et al, 2023) as well as the IWSLT 2022 human evaluation campaign (Anastasopoulos et al., 2022). It is suggested that the scores of the evaluation are stabilized when using such guidelines. It should be stated here that the evaluation here do not include giving a 1-100 score as it was done in WMT. The annotating team in this study simply ticks a score that falls between 0 and 6.

It should be noted here that LLM evaluation metrics have been introduced in the literature, such as the one introduced in Kocmi and Federmann (2023). It is, however, pointed out that LLMs favor translation outputs that are produced by the LLMs themselves (Liu et al., 2023). Accordingly, the researchers in this study exclude this type of evaluation since a NMT system is involved to avoid evaluation bias.

The evaluation in this study was performed by the professional annotators as mentioned earlier. The evaluation results are shown in **Table 1** below.

Table 1 above shows that Google Translate achieves an average score of 3.93 with a standard deviation of 0.23, and that the average score achieved by ChatGPT is 3.22 with a standard deviation of 0.19.

3.3 Error classification

The classification of errors in the study follows the error typology provided by Multidimensional Quality Metrics (MQM), introduced in Lommel et al. (2014). The typology provided by MQM classified translation errors into eight dimensions: terminology, accuracy (adequacy), linguistic conventions (fluency), style, locale conventions, audience appropriateness, design and markup, and custom. Such dimensions are defined and classified further. For instance, the main issues of accuracy are mistranslation, addition and omission, whereas the main issues related to fluency are grammar, punctuation and spelling.

In this study, the researchers used the MQM framework since it is viable, well-established and flexible. Numerous studies have used its typology to classify translation errors in the literature of translation research.

Since the focus is on text translation, annotators are asked to annotate intrasentential and intersentential errors. By identifying intersentential errors such as intersentential agreement, intersentential cohesion and lexical inconsistency, the analysis provides a better insight on the document-level rather than the sentence-level translation quality.

The annotated errors in the translated texts are categorized under 20 types as presented below. An explanation of the errors are provided with examples from the translated texts.

Terminology: Errors under this dimension occur when a term in the target text does not correctly represent the one in the source text where it can be incorrect, inconsistent, or inaccurate. The errors annotated in the texts under investigation that fall under this dimension are classified into the following categories:

- **Inconsistent Use of Terminology:** This category of errors refers to cases where multiple terms are used in the target text to represent the same term where consistency is required. An example of this error is the translation of the English term ”function” as both "دالة" and "وظيفة" in the same text.
- **Wrong Term:** This type of error refers to the use of a term that a professional translator will not use in a certain context or the use of a term that can cause

Table 1. Results of annotators' evaluation.

MT system	Mean	Standard deviation
Google Translate	3.93	0.23
ChatGPT	3.22	0.19

conceptual mismatch. An example of this error is the translation of the English term "abbreviated" as "تخفيض" where "اختصار" is more correct and suit- able to the context.

Accuracy: This refers to errors that arise when the content of the target text does not accurately match the propositional content of the source text. This can be for various reasons such as distortions, omissions, or additions to the message. The annotated errors in the texts under study that fall under this dimension follow the following subcategories.

- **Addition:** This refers to issues where an added content words/ phrases are included in the target text but not in the source text. An example of this is translating "augmentation of angiotensin" as " إفراز الألدوستيرون المتوسط" where the word "المتوسط" has no equivalent in the source text.
- **Ambiguous Target Content:** This refers to the case in which the target text or a part of it is ambiguous, i.e., it can be potentially interpreted in more than one way. The Arabic word "فطري" in Arabic can be understood as "inborn" or "fungous".
- **Ambiguous Source Content:** This refers to the case in which the source text or a part of it is ambiguous, i.e., it can be potentially interpreted in more than one way. The word "function", for example, is considered to be ambiguous and it can be translated into different Arabic words with completely different meanings.
- **Overly Literal:** This refers to the word for word translation when an idiomatic translation is required. Translating the sentence "Viruses take this notion to the extreme." into " إلى الحد الأقصى تأخذ الفيروسات هذا المفهوم. " is a clear example of this.
- **Omission:** This points to not translating a content that is present in the source text. A sentence like " يشير إلى عدد العناصر التي يمكن وضعها في المصفوفة." seems to be missing the subject.
- **Overtranslation:** This refers to an issue where the translation is more specific than the source text. Translating the word "parents" into "آباء" which means "fathers" where "والدين" is more suitable is an example of this issue.
- **Hallucination:** This refers to a completely different meaning from the source text created by machine translation. For example, " الحمل على الجزء الأيسر من القلب إلى زيادة التوتر على الجدار للجدار البطيني مع نتائج تشوه وظيفة الانبساط يؤدي فرط الأيسر للبطين، مما يحفز النمو الهاجس " as a translation for "There is an increase in total blood volume in proportion to body weight resulting in higher cardiac output" is really a hallucination in the word "الهاجس".
- **Undertranslation:** Translating "left ventricle" into "الجزء الأيسر من القلب" is an example of this issue where a general translation is given to a more specific term in the source text.
- **Untranslated:** In the current study, this type of error pertains to a segment of the target text that was supposed to be translated but it is not. An acronym like ISA, for example, is left as it is in the target text. Another example is translating the word "binary" as "باينارية" which is not used like this in scientific texts. However, when borrowing such terms is not the norm, the translation is considered to be incorrect and an error is annotated.

Linguistic Conventions (Fluency): Errors that belong to this dimension are related to the well-formedness of the text from a linguistic point of view. Under this dimension, the annotated errors in the texts under study are classified further into the following:

- **Word Form:** This type of error refers to the case where an inappropriate morphological variant of a word is chosen. This includes tense, agreement, part of speech, etc. An example of this error is the translation of the word "is located" as "يقع" instead of "تقع" in a context where it refers to "the nucleus" (النواة).This translation shows a mismatch in gender. The word "nucleus" is neuter in English, however its equivalent in Arabic is considered as feminine.
- **Word Order:** This simply refers to the incorrect

word order of the translated text which may follow the rules of the source language rather than those of the target language. An Arabic translation like "هياكل البيانات التسلسلية لديها العديد من الاستخدامات ". would be much better if it was "البيانات التسلسلية العديد من الاستخدامات لدى هياكل"

- **Incorrect Function Word:** This is related to the error of using incorrect function word, such as an incorrect article or an incorrect preposition. For example, ”they” has been translated as "هم" instead of "هي" in "Viruses are important to biologists for several reasons. They are the simplest form of life". The translation is " هم أبسط أشكال الحياة الفيروسات مهمة لعلماء الأحياء لعدة أسباب." Since "الفيروسات" are feminine, it should have been translated as "هي".
- **Missing Function Word:** This type of error refers to the case where a function word is required but it is not present in the target text. In the translation of ”because life is so diverse”, the adverb ”so” is missing in the translation "نظراً لتنوع الحياة". It should have been translated as " في الحياة/ نظراً لأن الحياة شديدة التنوع نظراً للتنوع الشديد".
- **Extraneous Function Word:** As opposed to the ”missing function word” error, this error refers to the existence of unnecessary function word in the target text. The first preposition "ب" in "بالثنائي" in the translation تُعرف التعليمات المشفرة بالثنائي باسم رمز الآلة as a translation for ”Binary-encoded instructions are known as machine code” is an obvious example of extraneous function word regardless of the agreement which is also required in the translation.
- **Punctuation:** This refers to the use of punctuation marks that are considered to be incorrect based on the rules of the target language. The commas which are used in وما إلى ذلك تتنوع الأعراض ولكنها قد تشمل صعوبة والدوخة، والخفقان، والأوجاع والآلام، والتعب، وضعف التركيز، في التنفس، والإغماء، as a translation for "Symptoms are variable but may include difficulty breathing, faintness, dizziness, palpitations, aches and pains, fatigue, impaired concentration and so on" are not necessary in Arabic since the conjunction "و" should be added to each coordinated noun (phrase) in the list as opposed to English where commas separate items in the list and the conjunction ‘and’ occurs after the penultimate item.
- **Spelling:** Errors related to miswriting words are included under this category. In the translation of ”The problem then is that every processor design would have a different architecture ……”, the word "`إذن" instead of "إذاً" is created in the translation " مختلفة المشكلة إذن هي أن كل تصميم للمعالج سيكون له بنية....”.
- **Cohesion:** Cohesion errors refer to a missing or incorrect part of a text that must be connected into a comprehensible whole. In the following two translated sentences " يتم تحديد طول أو حجم المصفوفة من العناصر التي يمكن وضعها في المصفوفة "خلال حجم كتلة الذاكرة مقسومًا على حجم العنصر. يشير إلى عدد there is no cohesion since the second sentence should start, for example, with "ويشير ذلك إلى عدد...." as a translation for " The length, or size, of an array is determined by the size of the memory block divided by the element size. It indicates how many elements can fit into the array."

Style: The target text here is grammatical and any error included under this dimension are grammatically acceptable. These errors exhibit inappropriate or awkward language style or they deviate from target language register. The translation of ”the genetic basis for inheritance in organisms” as " في الكائنات الأساسي الوراثي للتوريث " shows obvious deviation from the target language register where "الوراثي للتوريث" could have been replaced by "الجيني للوراثة" since the text is about biology.

Table 2 below presents the number and type of errors annotated in the translated texts from English to Arabic using Google Translate and ChatGPT based on MQM taxonomy.

3.4 Error frequency and distribution

It is mentioned earlier that the annotated texts are composed of 33 texts that contain 209 sentences. To calculate the error frequency in the produced translation, the number of total errors is divided by the number of the sentences in the translated texts. Accordingly, the error frequency is 1.31 per sentence in the Google Translate translation outputs and 2.02 in the ChatGPT translation outputs.

According to MQM main dimensions, the distribution of errors in Google Translate is as follows: 70 terminology errors (25.18% of the annotated errors), 42 accuracy errors

Table 2. Number of Errors in the translated texts in Google Translate and ChatGPT.

Dimensions	Types of errors	Google translate	ChatGPT
Terminology	Wrong Term	61	118
	Inconsistent use of terminology	9	1
Total of Terminology Errors		**70**	**119**
Accuracy	Ambiguous target content	1	1
	Ambiguous source content	2	2
	Overly literal	12	16
	Untranslated	4	15
	Undertranslation	1	2
	Overtranslation	1	0
	Hallucination	0	1
	Addition	0	8
	Omission	21	5
Total of Accuracy Errors		**42**	**50**
Linguistic conventions (Fluency)	Word Form	29	47
	Word Order	11	27
	Incorrect FW	10	8
	Missing FW	5	19
	Extraneous FW	11	17
	Punctuation	8	10
	Spelling	1	1
	Cohesion	57	64
Total of Linguistic conventions (Fluency) Errors		**132**	**193**
Style		34	60
Total of Style Errors		**34**	**60**
TOTAL ERRORS		**278**	**422**

(15.11% of the annotated errors), 132 fluency errors (47.48% of the annotated errors), and 34 style errors (12.23% of the annotated errors).

In ChatGPT translated texts, the distribution of errors are as follows: 119 terminology errors (28.2% of the annotated errors), 50 accuracy errors (11.85% of the annotated errors), 193 fluency errors (45.73% of the annotated errors), and 60 style errors (14.22% of the annotated errors).

The distribution of errors according to MQM main dimensions is represented in **Figure 1** below.

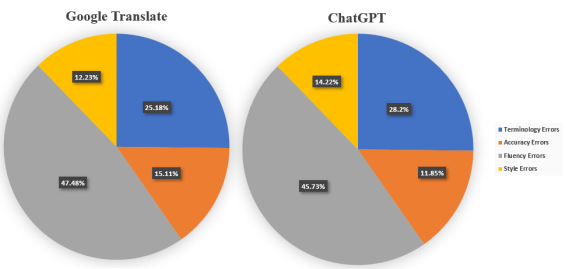

Figure 1. Error Distribution in Google Translate and ChatGPT.

4. Discussion

The results of the study show that there is a variety of annotated errors in the translation outputs of both Google Translate and ChatGPT. Both translation outputs vary in regard to the number of annotated errors. These errors also show variation depending on the main MQM dimensions in which they are included.

In general, Google Translate showed fewer errors than ChatGPT which is in accord with the professional evaluation of the translated texts. This means that Google Translate performs better than ChatGPT when translating English scientific texts into Arabic. However, this does not mean that the translation outputs of Google Translate are optimal. Both the evaluation and the number of annotated errors in the texts translated by Google Translate showed that it still falls short when translating scientific texts from English into Arabic. The results above show that the total number of annotated

errors in the translation outputs of Google Translate are 278 compared to 422 annotated errors in the translation outputs of ChatGPT. Similarly the evaluation results show that Google Translate achieves a higher average score of 3.93 compared to the 3.22 average score achieved by ChatGPT.

The results also show thar there are a lot of terminology errors in the translation outputs of both Google Translate and ChatGPT, which denotes that both systems fail greatly in that aspect. This also identifies that both systems need more training in regard to scientific text translation. It also shows the need of creating of more specialized annotated Arabic corpora. Nonetheless, it should be mentioned that Google Translate has performed better in this area considering the number of detected errors. The total number of annotated terminology errors in the translation outputs of Google Translate are 70 errors, whereas total number of annotated terminology errors in the translation outputs of ChatGPT are 119 errors.

Moreover, fluency errors showed the highest percentage among the annotated errors in the translation outputs of both Google Translate and ChatGPT. The results show that the annotated fluency errors account for 47.48% of the annotated errors in the translation outputs of Google Translate which was the highest as shown in Section 3.4. Similarly, the annotated fluency errors account for the highest percentage with 45.73% of the annotated errors in the translation outputs of ChatGPT. This denotes that both systems struggle to grasp the structural variations between English and Arabic. ChatGPT suffers much more in this aspect.

Arabic has a highly syndetic discourse (Farghal, 2017), and accordingly it uses more conjunctions than English. Conjunctions like "wa-" (and) and "fa-" are used at the beginning of a sentence to affirm its cohesion with the preceding contexts. If the conjunctions are not used, a text will seem rather unnatural. Therefore, the high number of cohesion errors (57 in Google Translate and 64 in ChatGPT) indicates that these systems still do not have a good grasp of this variation between English and Arabic and that more training for both systems is required in this aspect.

Along with cohesion, word form errors account for a high percentage of the annotated fluency errors. This shows that the systems under study face a big challenge when translating from a language with poor morphology (English) to a language with rich morphology (Arabic).

5. Conclusion

To conclude, it can be stated that Google Translate performed better than ChatGPT when translating English scientific texts into Arabic. However, this point toward the conclusion that both systems still fall short in this aspect and still need a lot of investigation and fine-grained analyses of the nature of errors in different types of texts. The need of training MT system, especially ChatGPT, is also crucial. More annotated corpora need to be built as well.

This study and other studies of the same nature are required since they provide invaluable insights on the strength and weakness of MT systems. The study is beneficial for translators, researchers and developers of machine translations. Therefore, more fine-grained studies on various text genera are recommended.

Author Contributions

The authors have contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

Funding

This research received grant no. (61/2023) from the Arab Observatory for Translation (an affiliate of ALECSO), which is supported by the Literature, Publishing & Translation Commission in Saudi Arabia.

References

Aghai, M., 2024. ChatGPT vs. Google Translate: Comparative analysis of translation quality. Iranian Journal of Translation Studies. 22(85). Available from: https://journal.translationstudies.ir/ts/article/view/1156

Ahmadnia, B., Dorr, B.J., 2020. Low-resource multi-domain machine translation for Spanish-Farsi: Neural or statistical? Procedia Computer Science. 177, 575–580.

DOI: https://doi.org/10.1016/j.procs.2020.10.081

Almekhlafi, H.A., Nagi, K.A., 2024. Fine-grained evaluation of English to Arabic neural machine translation: A case study of education research abstract. Al-Andalus Journal for Humanities & Social

Sciences, 95(11).

DOI: https://doi.org/10.35781/1637- 000-095-007

Barrault, L., Bojar, O., Costajussà, M. R., et al., 2019. Findings of the 2019 Conference on Machine Translation (WMT19). Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). DOI: https://doi.org/10.18653/v1/w19-530

Bentivogli, L., Bisazza, A., Cettolo, M., et al., 2016. Neural versus phrase-based machine translation quality: A case study. arXiv preprint arXiv:1608.04631.

DOI: https://doi.org/10.48550/arXiv.1608.04631

Bubeck, S., Chandrasekaran, V., Eldan, R., et al., 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

DOI: https://doi.org/10.48550/arXiv.2303.12712

Buscemi, A., Proverbio, D., 2024. Chat- GPT vs Gemini vs Llama on multilingual sentiment analysis. arXiv preprint arXiv:2402.01715.

DOI: https://doi.org/10.48550/arXiv.2402.01715

Castilho, S., 2020. On the same page? Comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In Proceedings of the Fifth Conference on Machine Translation. pp. 1150–1159. Available from: https://aclanthol- ogy.org/2020.wmt-1.137

Castilho, S., Doherty, S., Gaspari, F., et al., 2018. Approaches to human and machine translation quality assessment. In: J. Moorkens, S. Castilho, F. Gaspari., et al. (Eds.). Machine Translation: Technologies and Applications. Springer International Publishing.

DOI: https://doi.org/10.1007/978-3-319-91241-7

Chan, C., Jiayang, C., Wang, W., et al., 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In: Findings of the Association for Computational Linguistics: EACL 2024. pp. 684–721. Available from: https://aclanthology.org/2024.findings-eacl.47

Daems, J., Macken, L., Vandepitte, S., 2014. On the origin of errors: a fine-grained analysis of MT and PE errors and their relationship. In 9th International Conference on Language Resources and Evaluation (LREC). pp. 62–66. European Language Resources Association (ELRA). Available from: http://hdl.handle.net/1854/LU- 4418636

Escartín, C.P., Goulet, M.-J., 2020. When the post-editor is not a translator. Can machine translation be post-edited by academics to prepare their publications in English? In: Translation Revision and Post-Editing. Routledge. pp. 89–106.

Farghal, M., 2017. Textual issues relating to cohesion and coherence in Arabic/English translation. Jordan Journal of Modern Languages and Literature, 9(1), 29–50.

Freitag, M., Foster, G., Grangier, D., et al., 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. Transactions of the Association for Computational Linguistics, 9, 1460–1474.

DOI: https://doi.org/10.1162/tacl_a_00437

Graham, Y., Haddow, B., Koehn, P., 2019. Translationese in machine translation evaluation. arXiv e-prints, arXiv-1906.

DOI: https://doi.org/10.48550/arXiv.1906.09833

Graham, Y., Haddow, B., Koehn. P., 2020. Statistical power and translationese in machine translation evaluation. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

DOI: https://doi.org/10.18653/v1/2020.emnlp-main.6

Hassan, H., Aue, A., Chen, C., et al., 2018. Achieving human parity on automatic Chinese to English news translation. arXiv preprint arXiv:1803.05567.

DOI: https://doi.org/10.48550/arXiv.1803.05567

Hendy, A., Abdelrehim, M., Sharaf, A., et al., 2023. How good are GPT models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210.

DOI: https://doi.org/10.48550/arXiv.2302.09210

Hicks, M.T., Humphries, J., Slater, J., 2024. ChatGPT is bullshit. Ethics and Information Technol- ogy. 26(2), 38.

DOI: https://doi.org/10.1007/s10676-024-09775-5

Isabelle, P., Cherry, C., Foster. G., 2017. A challenge set approach to evaluating machine translation. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

DOI: https://doi.org/10.18653/v1/d17-1263

Jiao, W., Wang, W., Huang, J.T., et al., 2023. Is ChatGPT a good translator? A preliminary study. arXiv preprint arXiv:2301.08745, 1(10).

DOI: https://doi.org/10.48550/arXiv.2301.08745

Khoshafah. F., 2023. ChatGPT for Arabic-English translation: Evaluating the accuracy.

Kocmi, T., Federmann, C., 2023. Large language models are state-of-the-art evaluators of translation quality. In: Proceedings of the 24th Annual Conference of the European Association for Machine Translation. pp. 193–203. https://aclanthology.org/2023.eamt1.19

Kocmi, T., Bawden, R., Bojar, O., et al., 2022. Findings of

the 2022 Conference on Machine Translation (WMT22). In Proceedings of the Seventh Conference on Machine Translation (WMT). pp. 1–45. Available from: https://aclanthology.org/2022.wmt-1.1

Koehn, P., Knowles. R., 2017. Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation.

DOI: https://doi.org/10.18653/v1/w17-3204

Läubli, S., Castilho, S., Neubig, G., et al., 2020. A Set of Recommendations for Assessing Human–Machine Parity in Language Translation. Journal of Artificial Intelligence Research. 67.

DOI: https://doi.org/10.1613/jair.1.11371

Läubli, S., Sennrich, R., Volk. M., 2018. Has machine translation achieved human parity? A case for document-level evaluation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

DOI: https://doi.org/10.18653/v1/d18-1512

Levin, P., Dhanuka, N., Khalilov, M., 2017. Machine translation at booking.com: Journey and lessons learned. arXiv preprint arXiv:1707.07911.

Liu, J., Liu, C., Zhou, P., et al., 2023. Is ChatGPT a good recommender? A preliminary study. arXiv preprint arXiv:2304.10149.

DOI: https://doi.org/10.48550/arXiv.2304.10149

Lommel, A., Uszkoreit, H., Bur- chardt, A., 2014. Multidimensional quality met- rics (MQM): A framework for declaring and describing translation quality metrics. Tradumàtica Tecnologies de La Traducció. 12, 455–463. DOI: https://doi.org/10.5565/rev/tradumatica.77

Nagi, K.A., 2023. Arabic and English relative clauses and machine translation challenges. Journal of Social Studies. 29(3), 145–165.

DOI: https://doi.org/10.20428/jss.v29i3.2180

Nagi, K.A., Alzain, E., Naji, E., 2024. Informed prompts and improving ChatGPT English to Arabic translation. Al-Andalus Journal for Humanities & Social Sciences. 98(11). Available from: https://www.researchgate.net/publication/382295323_Informed_Prompts_and_Improving_ChatGPT_English_to_Arabic_Translation

OpenAI, R., 2023. GPT-4 technical report. arxiv 2303.08774. 2, 13.

DOI: https://doi.org/10.48550/arXiv.2303.08774

Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., et al., 2023. Linguistic ambiguity analysis in ChatGPT. arXiv preprint arXiv:2302.06426.

DOI: https://doi.org/10.48550/arXiv.2302.06426

Poibeau, T., 2022. On “human parity” and “super human performance” in machine translation evaluation. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 6018–6023. Available from: https://hal.science/hal-03738720/

Popel, M., Tomkova, M., Tomek, J., et al., 2020. Transforming machine translation: a deep learn- ing system reaches news translation quality comparable to human professionals. Nature Communications. 11(1).

Popović, M., 2021. On nature and causes of observed MT errors. In Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track. pp. 163–175. Available from: https://aclanthology.org/2021.mtsummit-research.14

Reeder, F., 2004. Investigation of intelligibility judgments. In: Conference of the Association for Machine Translation in the Americas. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 227–235.

DOI: https://doi.org/10.1007/978-3-540-30194-3_25

Rivera-Trigueros, I., 2021. Ma- chine translation systems and quality assessment: a systematic review. Language Resources and Evaluation. 56(2), 593–619.

Saunders, D., 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. Journal of Artificial Intelligence Research. 75, 351–424.

Sennrich, R., Zhang. B., 2019. Revisiting low-resource neural machine translation: A case study. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

DOI: https://doi.org/10.18653/v1/p19-1021

Siu. S. C., 2023. ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models in translation. SSRN Electronic Journal.

DOI: https://doi.org/10.2139/ssrn.4448091

Tehseen, I., Tahir, G.R., Shakeel, K., et al., 2018. Corpus based machine translation for scientific text. In L. Iliadis, I. Maglogiannis., V. Plagianakos (Eds.), Artificial Intelligence Applications and Innovations. Springer International Publishing. pp. 196–206.

DOI: https://doi.org/10.1007/978-3-319-92007-8

Toral, A., 2020. Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. arXiv preprint arXiv:2005.05738.

DOI: https://doi.org/10.48550/arXiv.2005.05738

Toral, A., Sánchez-Cartagena, V. M., 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers.

Toral, A., Castilho, S., Hu, K., et al., 2018. Attaining the unattainable? Reassessing claims of human parity in neural machine translation. Proceedings of the Third Conference on Machine Translation: Research Papers.

DOI: https://doi.org/10.18653/v1/w18-6312

Ulitkin, I., Filippova, I., Ivanova, N., et al., 2021. Automatic evaluation of the quality of machine translation of a scientific text: The results of a five-year-long experiment. E3S Web of Conferences. 284, 08001.

DOI: https://doi.org/10.1051/e3sconf/202128408001

Zhu, W., Liu, H., Dong, Q., et al., 2023. Multilingual machine translation with large language models: Empirical results and analysis. arXiv e-prints, arXiv-2304.

DOI: https://doi.org/10.48550/arXiv.2304.04675

Zulfiqar, S., Wahab, M.F., Sarwar, M.I., et al., 2018. Is machine translation a reliable tool for reading German scientific databases and research articles? Journal of Chemical Information and Modeling. 58(11), 2214–2223