Assessing the Quality of ChatGPT-Generating Arabic Subtitles: A Functional and Linguistic Analysis

Ben-milad, Khaled

English Department, El-mergib university, Khomes, Libya

kmbenmilad@elmergib.edu.ly

Abstract

The emergence of generative artificial intelligence has reshaped the translation industry. Large language models can do more than just translate; generate, edit, and review texts. However, little is known about auto-generated subtitling. Effective subtitling requires deep technical, linguistic and cultural knowledge. This study seeks to evaluate automatic subtitling produced by the OpenAI's ChatGPT-3.5, focusing on linguistics and functional equivalence. Using the content of TED talks: the original transcript as the source material, Arabic subtitles as reference translation, and the platform subtitle guidelines as prompts, the experiment was conducted. The automatically generated subtitles were analysed by human translators according to the parameters of FAR: functional equivalence, readability and acceptability. The findings show that omission is the most frequent error category in the auto-generated subtitling, affecting the intended audio message and informational details. Less frequent issues were found such as semantic, and stylistic errors, compressed sentences, and mistranslation. This suggests that auto-generated subtitling is generally understandable but still lacks accuracy. Further research is needed to investigate the advanced ChatGPT's auto-generated subtitling.

Keywords: generative artificial intelligence, auto-generated subtitling, ChatGPT 3.5, TED talk, English-to-Arabic translation.

الملخص

أدى ظهور أدوات الذكاء الاصطناعي التوليدي إلى إعادة تشكيل صناعة الترجمة، حيث لم تعد النماذج اللغوية الضخمة تقتصر على الترجمة فقط، بل أصبحت قادرة على توليد النصوص وتحريرها ومراجعتها. غير أن المعلومات المتوفرة حول قدرتها في مجال الترجمة المرئية لا تزال محدودة جدًا، رغم أن هذا النوع من الترجمة يتطلب معرفة تقنية ولغوية وثقافية عميقة لضمان فعاليته. تهدف هذه الدراسة إلى تقييم

الترجمة المرئية الناتجة عن استخدام شات جي بي تي بوصفه أحد نماذج الذكاء الاصطناعي التوليدي، مع التركيز على الترجمة المرئية إلى اللغة العربية. وقد أُجريت الدراسة بالاعتماد على محتوى محادثات من منصة "تيد"، حيث استُخدمت النسخة المفرغة صوتيًا كنص مصدر، والترجمة التوضيحية كنموذج مرجعي، إضافة إلى إرشادات الترجمة بصفتها هندسة أوامر موجهة لشات جي بي تي. وقد تم تحليل وتقييم الترجمات التوضيحية المولدة آليًا من قبل مترجمين متخصصين، وفقًا لمعايير المطابقة الوظيفية، وسهولة القراءة، وقبول المشاهد. وأظهرت النتائج أن الحذف يُعد أكثر أنواع الأخطاء شيوعًا في الترجمة المرئية المولدة آليًا، مما قد يؤثر على المعنى المقصود والتفاصيل الدقيقة للنص. كما شملت الأخطاء الأخرى، وإن كانت أقل تكرارًا، الأخطاء الدلالية والأسلوبية، بالإضافة إلى حالات من الترجمة المقتضبة التي أخلّت بالمعنى، وأخطاء ترجمة مباشرة غير دقيقة. وتخلص الدراسة إلى أن الترجمات المرئية المولدة آليًا تُعد مفهومة بشكل عام، إلا أنها لا تزال تفتقر إلى الدقة المطلوبة. وعليه، توصي الدراسة بضرورة إجراء المزيد من الأبحاث لاستكشاف إمكانات النماذج المتقدمة من شات جي بي تي في مجال الترجمة المرئية وتحسين أدائها

الكلمات المفتاحية :الذكاء الاصطناعي التوليدي الترجمة المرئية، شات جي بي تي ,تيد والترجمة من الإنجليزية إلى العربية

Introduction

Since artificial intelligence (AI) emerged, automated translation has grown to play a leading role across many areas of the translation industry. Among the different types of AI is neural machine translation (NMT), it uses deep learning mechanisms to predict word sequences and translate entire sentences, enabling an increasingly accurate automation translation when compared with that of previous MT methods (Bahdanau et al., 2015; Sennrich et al., 2016; Bojar et al., 2016). The rise of NMT quality has led to the expansion of automation in audiovisual translation (AVT) modalities such as subtitling (Castilho et al., 2017), however; acceptance within the subtitling community has been limited (Bywood et al., 2017). Matusov et al. (2019) give a reason for the lower rate of NMT adoption is that the systems do not fully address the specific issues and challenges involved in the subtitling process**.**

In late 2022, OpenAI released ChatGPT, serving a key milestone in the emergence of generative artificial intelligence (GenAI) (Marr, 2025). The

generative AI tools showed the ability to generate texts very similar to those created by humans, especially for high resource languages such as English. It says that is due to the large-scale training data, and the advantages of engineering prompts, so they help predict the next token in a sequence. The user can also ask to take context of source and target into consideration (Moorkens et al. 2025).

The audio-visual (AVT) industry is thus becoming just the latest whose markets can be expanded through automatic translation. AVT is experiencing exponential growth in the implementation of MT and GenAI solutions (Slator 2024). Subtitling, one of the primary modes of AVT, involves converting spoken language into written text on-screen, enabling audiences to access content across language barriers (Díaz-Cintas, 2003).
Subtitling can be challenging because of linguistic nuances and cultural differences. It is also constrained by technical elements and requires a high level of precision to preserve the original meaning. According to Díaz-Cintas & Remael (2021), the subtitling landscape is a multifaceted translation practice that requires technical, linguistic and cultural skills. Automatic subtitling adds further complexity, involving a multi-step process that includes automatic segmentation, transcription, spotting, speech recognition, and machine translation (Xie, 2023). This shift has led to increased demand for automatic subtitling in streaming platforms with low cost and quick development. The prominent streaming platform of YouTube for example, has already adopted machine-translated subtitling in multiple, diverse languages using automatic speech recognition and MT. In Al-Momani et al. study, (2025) found that while YouTube automatic subtitling can be useful, it often fails to convey meaning effectively into Arabic, negatively impacting comprehension and overall enjoyment. Participants showed a preference for human-created subtitles. Similarly, Al-Harahsheh & Rababah (2025) reported that automatic subtitling exhibited limitations in accurately conveying meaning.
ChatGPT has demonstrated ability to process machine translation including translation prompt, multilingual translation, and translation robustness (Jiao et al., 2023), which highlights the model's advanced natural language processing capabilities. However, the questions remain about its accuracy, and overall effectiveness compared to human translation. While ChatGPT

has numerous advantages, it also has several limitations, including accuracy issues, a lack of human touch, and privacy concerns (Yang et al., 2024). These challenges open new avenues for research in particular when the target is a highly rich morphology language such Arabic. Arabic subtitling presents unique challenges due to its morphological complexity, insensitivity to cultural aspects, and technical specificity (Al Sawi & Allam, 2024).

The current study aims to investigate the accuracy of automatic subtitling into Arabic generated by OpenAI ChatGPT. The analysis focuses on identifying any semantic and stylistic issues between the source text and the machine- translated subtitling in terms of their ability to accurately convey the intended meaning. This study also aims to analyse the subtitle readability (i.e. how easily the audience can read, process, and understand on-screen text), and acceptability (i.e. how natural and grammatically correct the subtitles sound to native speakers). Additionally, the objective of this study is to emphasise the importance of translator's role in subtitling compared to those produced by machine. Therefore, the study attempts to answer the following questions:

1. To what extent can auto-generated subtitling effectively convey accurate meaning, ensuring appropriate readability and perfect acceptability?
2. To what extent can ChatGPT respond to prompts in producing subtitles?
3. How is the quality of ChatGPT-generated Arabic subtitling compared to those created by humans?

The significance of the study arises from lack of experimental research that has explored the quality of auto-generated subtitling into Arabic. Quality subtitling determines whether a conversation is truly accurate, readable and accessible.

The remainder of this paper is organized as follows: Section 2 reviews related work on automatic subtitling. Section 3 describes the experimental setup. Section 4 presents the findings, and discussion of the results in Section 5, Finally, the conclusions drawn from the study are in Section 6.

Literature Review

Automatic subtitling has taken the stand in research trying to boost the translation industry. The quality of MT subtitling has been a focus of research. One direction emphasises on the good quality of machine-translated subtitling. Belles- Calvera and Caro Quintana (2021) demonstrate through using error typology analysis and automatic metrics, the quality of the English subtitle translation of the Spanish Netflix series Cable Girls provided by two NMT systems is good. Hiraoka and Yamada's study (2019) assesses automatic subtitling quality for TED presentations from Japanese to English by employing a combination of human raters and BLEU metric. Martinez and Vela (2016) carry out an analysis of errors to subtitling produced by humans to those created by the machine from English into German. They point out that although qualifying as acceptable, revisable MT subtitling outnumber human-translated segments. Hagstrom and Pedersen (2022), in contrast, conducted a diachronic study of subtitling before and after involving MT, by comparing a corpus of subtitles for Swedish national TV subtitling guidelines in the 2020s and before. The post-edit process, based on the FAR model, demonstrates a lower quality of subtitling since the 2020s, which they attribute to the increased use of MT.

Other research direction focuses on subtitlers' feedback on editing machine-translated subtitling. Xie (2023) examined post-editing efforts in automatic subtitling, the analysis of the user experience questionnaire shows that although the output suffer from some issues like mistranslation, wrong tense, literal translation, most subtitlers give a positive attitude towards the machine-translated subtitling. Koponen et al. (2020) investigated productivity of the automatically generated subtitling, comparing subtitlers' effort when post-editing MT subtitling against translating from scratch, in four Scandinavian languages. The analysis of reviewers' data found that although MTPE was slightly faster, and to involve fewer keystrokes than subtitling from scratch, critical issues like segmentation and timing of subtitling were found in the automatic subtitling. In spite of that, a majority of subtitlers maintain a favourable perspective on MT subtitling. Karakanta et al., (2022,) extended Koponen's study in subtitlers' impressions on using automatic subtitling, by involving a large sample size, and

different languages. The reviewers' analysis found that, despite some limitations stemming from failures in auto-spotting and segmentation, automatic subtitling is seen rather positively and can be beneficial and potential, and help subtitlers save time and effort. Hu et al. (2020) investigated the perception and reception of machine-translated subtitling of MOOCs (Massive Open Online Courses), through combination of a questionnaire and eye-tracking method. The study showed no significant difference in reception on MT, PEMT and HT subtitling. Most annotator's view machine- translated subtitling positively.

Review subtitling into Arabic, research compares the quality of human versus machine-generated subtitling across various platforms, examining the specific strategies used and the obstacles encountered.

Al-Darabee et al., (2025) compared Netflix's subtitling with Google Translate's renditions of English scatological expressions. The results indicate that human subtitlers preserved a balanced tone and respected cultural sensitivities in intercultural communication, whereas machine-translated subtitling were often inaccurate, leading to issues with fluency and coherence. Similarly, Al Sammarraie, (2025) investigated the challenges and strategies involved in post-editing Google NMT output for the Arabic subtitles of the Netflix movie '*La La Land'*, focusing on linguistic fidelity, cultural adaptation, and technical constraints. The findings demonstrate that the most frequent strategy was paraphrasing (localization), followed by cultural substitution, both used to enhance the efficiency and quality for Arabic audience. In contrast, machine-translated subtitling relied heavily on literal translation. Both studies emphasise the significant role of human expertise in improving the automatic translation.

Farghal & Ammari (2025) investigated the performance of automated subtitling on three social media platforms: YouTube, Facebook, and Instagram. Qualitative and quantitative analysis indicate that while YouTube outperformed the other two platforms, all three have a common challenge: inaccurate translation of colloquial, idiomatic, dialectical, and discursive expressions.

Al-Momani at al., (2025) explored the effectiveness of YouTube auto-generated subtitling in rendering the cultural connotation and maintaining linguistic accuracy. The data analysis reported that while reviewers acknowledged that the auto-generated subtitling improves accessibility for global media, the output often contained mistranslations and inaccurate cultural expressions, resulting in confusion and reduced audience engagement. The

study highlighted that the human post-editing remains vital to maintaining accuracy and cultural sensitivity. Similarly, Al-Harahsheh & Rababah (2025) analysed the automatic YouTube's subtitles of 'Harry Potter and the Prisoner of Azkaban' into Arabic using the FAR model. The findings showed that the auto-generated subtitles contained semantic errors, grammatical mistakes, and cultural misinterpretations. Further, the study identified weakness on the speech recognition systems. The results also emphasised that while automatic subtitling enhance accessibility and speed, the human post-editing significantly increase the linguistic accuracy and naturalness of audio.

Regarding AI-generated subtitling, little research investigated the quality of chatbot generated subtitling, Calvo-Ferrer (2023) conducted a study on the reception of human and auto-generated subtitling. The study examined the challenges faced by ChatGPT-generated subtitling in comparison with those created by human translators in English- Spanish translations. The findings indicate that, despite significant advancements in AI-generated subtitling, human-translated subtitles continue to outperform them in terms of accuracy and contextual comprehension. They also note that the results may vary depending on the nature of the source content and the target language. Al Sawi and Allam (2024) explored challenges of using the free version of ChatGPT when subtitling movie allusions in Birdman, comparing it to the strategies employed by human translators. The findings revealed that while human translators used linguistic skills and culture sensitively in rendering culturally specific references (allusions), ChatGPT auto-generated subtitling relied primarily on literal translation, leading to potential misunderstandings in literary and culture-specific content. Finally, Abdelaal and Al Sawi (2025), Moneus and Sahari (2024) investigated the effectiveness of using ChatGPT and NMT systems, within legal texts. They found that automatic translation faced critical challenges in capturing legal terminology as accurately as a human translator, further emphasising the complementary nature of human and machine translation. To the researcher's knowledge, there remains a lack of experimental reserach assessing the quality of AI-generated subtitling. This study therefore focuses on evaluating the accuracy of meaning transfer, as well as readability and accessibility, with the aim of identifying potential improvements in automatic Arabic subtitling.

Experiment design

This study will employ a quantitative analysis to evaluate the results of ChatGPT's subtitling to three TED talks in English-to-Arabic translation, based on the FAR model. This approach will help identify any significant technical, linguistic, and cultural issues that may lead to misunderstanding of the audio.

Evaluation method

The evaluation approach is based on the error identification derived from the FAR model to assess the quality of auto-generated subtitles. The model developed by Jan Pedersen (2017), is designed to evaluate interlingual subtitles across three key parameters: functional equivalence, acceptability, and readability. Functional equivalence - including both semantic and stylistic aspects - measures how effectively the subtitle conveys the intended meaning. Acceptability focuses on grammatical accuracy, idiomatic usage, and spelling, assessing how well the target text aligns with linguistic norms. Readability, on the other hand, addresses technical issues such as synchronization, inappropriate length or speed, and punctuation errors, all of which may hinder comprehension.

The errors are classified according to their level of severity: minor, standard, or serious. Serious errors prevent viewers from understanding the content or significantly distort its meaning. Standard errors may affect subtitle readability or result in a slight loss of meaning. Minor errors are technically incorrect but do not hinder the audience's understanding.

The rationale for adopting the FAR model lies in its comprehensive framework for interlingual subtitle quality assessment. It enables evaluators to systematically recognise and classify translation errors, including how well source language texts are transformed into equivalent target language terms, and to determine whether these translations are both acceptable and easily understood by the audience.

Data and subtitling guidelines

The dataset of this study was sourced from TED TALK (ted.com), which is regarded as a reliable platform for the experiment. The research utilized

TED's knowledge-driven streaming platform (Technology, Entertainment, Design), to extract source text for input, obtain reference translations for comparison, and gather subtitles guidelines used as prompts for ChatGPT.

TED is a global, non-profit media platform dedicated to freely sharing ideas with a worldwide audience under the slogan "Ideas Worth Spreading" primarily through short and impactful video presentations. The platform operates in English and provides multilingual subtitles translated by volunteers following specific subtitles guidelines[1].

The TED platform provides a set of guidelines for subtitles that are designed to accurately convey meaning. These include a maximum of 42 characters per line (including spaces) and no more than two lines per subtitle. The length of the two lines should be kept as balanced as possible. When breaking lines, linguistic units should remain intact. Sentences should not be combined or split unless necessary. These guidelines are used as prompts for ChatGPT.

The dataset consists of three Talks with Arabic subtitles, covering diverse topics: history, culture, education. Both the original English transcripts and their corresponding Arabic subtitles were included in the corpus of study. The human-translated subtitles to those talks served as the gold standard reference during annotators' analysis of errors in the ChatGPT-generated output. Table 1 below presents the selected TED presentations, along with their segmentation counts, duration, and view frequencies.

[1] TED TALK Subtitling Tips

Table 1: TED TALKs, along with their segmentation counts, duration, and view frequencies.

TED talk topic	segmentation counts	Duration	frequency
What really happened to the Library of Alexandria?[2]	89	04:53	2.996,194
Why should you read Charles Dickens?[3]	87	5:17	2,048,467
THE HIDDEN TREASURES OF TIMBUKTU[4]	84	5:35	1.035,217

AI Translation Tools and prompts

ChatGPT 3.5 (https://chatgpt.com/), a widely used and freely available text-based generative AI tool, was selected for this study. This choice was made due to the preliminary nature of the research. In a subsequent phase, the study would be extended to examine the performance of more advanced ChatGPT models.

In the experiment, we submitted the prompts below to ChatGPT 3.5.

[2] Elizabeth Cox: What really happened to the Library of Alexandria? | TED Talk
[3] Iseult Gillespie: Why should you read Charles Dickens? | TED Talk
[4] ELIZABETH COX: THE HIDDEN TREASURES OF TIMBUKTU | TED TALK

Table 2: Prompt 1 on ChatGPT 3.5

Prompt: 1 You work as a subtitler for translating the English transcript that I will be providing, into Arabic. Please use the following guidelines for the output: There is a maximum of 42 characters per line (including blanks). Each subtitle has a maximum of two lines. Keep the two line-length balance close. Keep linguistic units together when breaking lines. Do not merge two sentences, or split sentences if not necessary. The subtitles are produced for a TED talk.

Table 3:: Prompt 2 on ChatGPT 3.5

Prompt 2: Please Translate the English transcript subtitles into Arabic subtitles, using guidelines indicated above.

Findings

The ChatGPT auto-generated subtitles were analysed using FAR error categories. Figure 1 below provides a summary of the error types and their frequencies within each category.

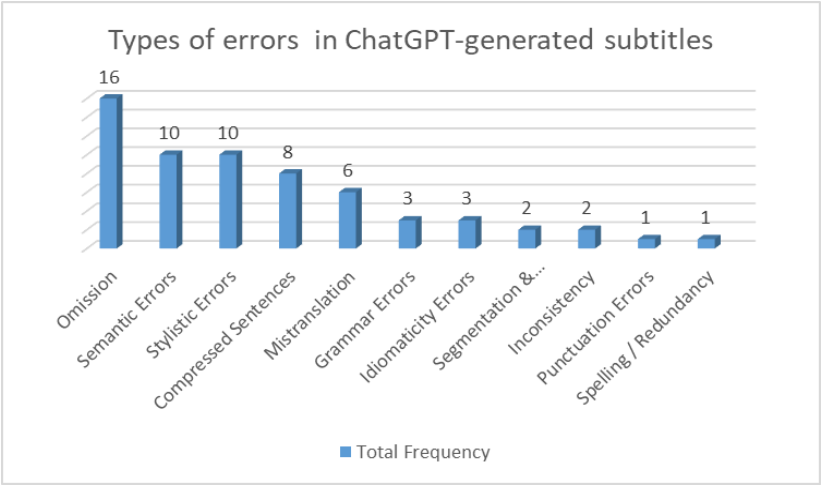


Figure 1: Distribution of error types

In the figure above, omission is identified as the most frequent error category, where elements such as modifiers and descriptive details present in the source are removed in the auto-generated subtitles.

For example,

ST: and still more remain lost or hidden in the desert

HT-subtitle: ويبقى المزيد منها مفقودًا أو مُخبأ في الصحراء

ChatGPT subtitle: وأخرى ما تزال مفقودة بالصحراء

In the example above, the auto-generative subtitling omits the descriptive unit of "hidden", which appears in the source text and translates accurately by the human translation, thereby reducing informational accuracy. This is classified as a minor error, as it does not significantly hinder the audience's understanding.

Semantic errors

A semantic error occurs when a subtitle is grammatically correct but conveys an incorrect meaning, thereby distorting the intended message of the audio.

For example,

ST: Christian rulers even had a mathematician murdered

HT-subtitle: كان لدى الحكام المسيحيين عالم رياضيات قُتل

ChatGPT subtitle: قتل الحكام المسيحيون العالِمة

The automatically generated subtitle alters the semantic nuance from indirect causation to direct killing, resulting in a shift in meaning. This is classified as a serious issue, as it changes the sentence's semantic meaning, affecting the viewer's understanding of the preceding context.

Stylistic errors

A stylistic error occurs when a subtitle technically conveys the intended message but does not align with the tone or register of the original dialogue.

For example,

ST: gave rise to its own adjective

HT-subtitle: أعطاها صفة خاصة بها

ChatGPT subtitle: أنجبت صفة خاصة بها

The automatically generated subtitle translates the phrase "gave rise" as 'أنجبت' in a way that makes the sentence sound less natural in Arabic. This is considered a minor error, as it does not significantly affect the audience's understanding.

Compressed sentences

Compressed sentences that retain the most important information while using fewer words.

For example,

ST: must negotiate with a shadowy cast of characters

HT-subtitle: يجب أن يتفاوض مع مجموعة من الشخصيات الغامضة

ChatGPT subtitle: ويتعامل مع شخصيات غامضة

The auto-generated subtitle simplifies the structure and removes clarifying details such as "a shadowy cast," which results in reduced clarity. This is considered a standard error because it impacts the meaning of the sentence rather than the overall scene.

Mistranslation

Mistranslation occurs when spoken dialogue is rendered inaccurately or misleadingly as on-screen text.

For example,

ST: boot blacking factory

HT-subtitle: مصنع للطلاء الأسود

ChatGPT subtitle: مصنع أحذية

The automatically generated subtitle incorrectly translates the descriptive term "blacking" (referring to shoe-polishing work) as "shoe factory," which changes the intended meaning. This is classified as a serious error, indicating a significant semantic issue in the output.

Grammar errors

Grammar mistakes involve inaccuracies in syntax and tense that hinder readability.

For example,

ST: hundreds of thousands of books

HT-subtitle: مئات الآلاف من الكتب

ChatGPT subtitle: مئات آلاف الكتب

The auto-generated subtitle contains an error: the omission of the definite article "الـ" in the word "آلاف", which results in a less natural structure. This issue is classified as a minor error.

Idiomaticity errors

Idiomatic errors occur when a translated subtitle does not follow the natural usage conventions of the target language.

For example,

ST: became the most sought-after commodity

HT-subtitle: أصبحت منتجات تلك الثقافة الفكرية الآن السلعة الأكثر رواجًا

ChatGPT subtitle: أصبحت منتجاتها الفكرية أغلى السلع طلبًا

The automatically generated phrase "أغلى السلع طلبًا" sounds unnatural in Arabic, whereas the human translation "السلعة الأكثر رواجًا" is more idiomatic and therefore more acceptable to native speakers. Although this results in some awkwardness in expression, it is classified as a minor error that does not alter the overall meaning.

No significant issues were identified in the output, including segmentation and spotting, punctuation, spelling, and line length, suggesting that ChatGPT adhered reasonably well to the guidelines.

Even so, it is important to recognise that automatically generated subtitles, while not yet achieving perfect quality, can be accurate in certain cases. An example is provided below,

ST: the past is obsolete

HT: الماضي هو شيء منتهي الصلاحية

ChatGPT: الماضي عفا عليه الزمن

In this example, ChatGPT's output shows stylistic improvement, using more natural Arabic expressions than those produced by the human translator.

Discussion

The analysis of the auto-generated subtitles reveals several recurring error categories with different levels of severity when analysed and compared with the human translation. Figure 2 below displays the error categories in percentage.

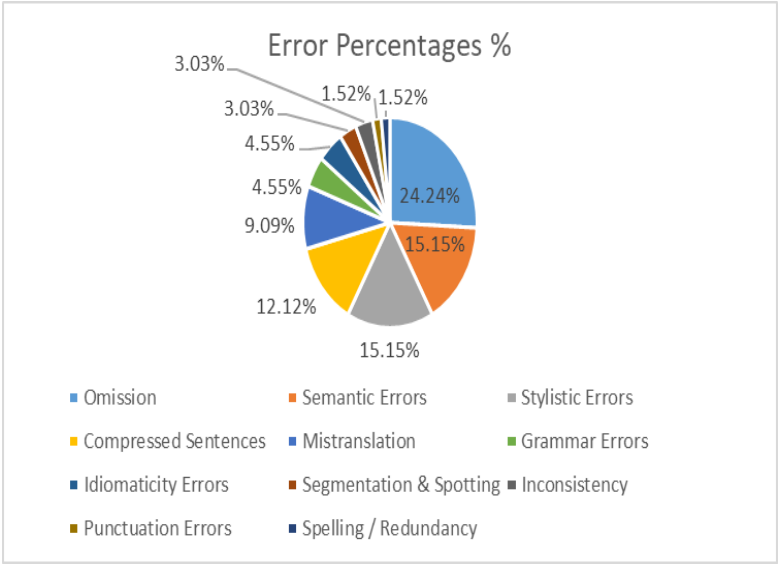


Figure 2: Percentage of error types

Figure 2 above demonstrates that omission is the most frequent error type, accounting for (24%) of the total errors. This suggests that ChatGPT 3.5 occasionally failed to convey certain lexical or semantic elements from the source into the target, potentially affecting the meaning and information details. The tendency to short segments, likely due to subtitle constraints such as limited space and reading speed, which may contribute significantly to meaning distortion. Semantic, and stylistic errors each representing (15%) of errors. These issues indicate that while ChatGPT preserves the general meaning, it slightly alters the semantic nuance of the original English text or produces less natural expressions in Arabic. Compressed sentences (12%) also reflect a tendency toward reduction, resulting in partial loss of information, and occasionally reducing clarity of the original message. Although compressed sentences are often necessary for subtitle brevity, excessive use of compression can lead to the loss of important semantic information. Less frequent issues such as mistranslation, representing around (9%), also contribute significantly to meaning distortion. Minor categories such as grammar errors (5%), idiomaticity errors (5%), and punctuation errors and spelling/redundancy each account for (2%), appeared less frequently but still affected the naturalness and subtitle readability.

Regarding severity of errors, Figure 3 below displays the levels of severity in percentage.

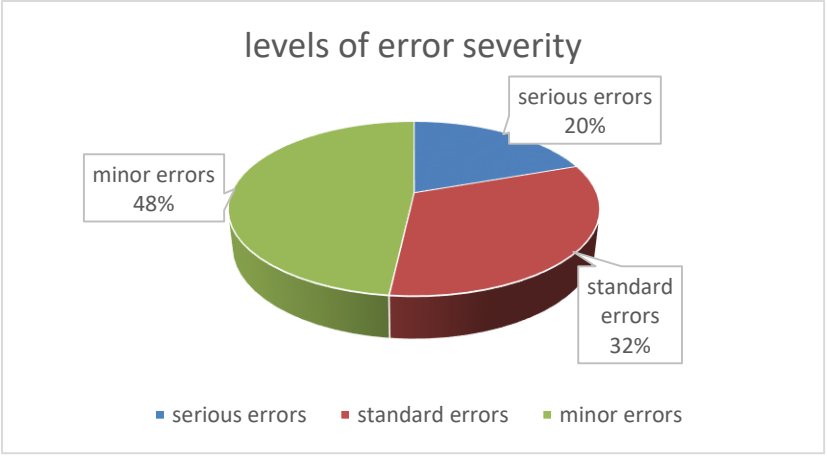


Figure 3: levels of severity

In Figure 3 above, minor errors dominate (48%), indicating the most AI generated-subtitles issues are no-critical. Standard errors (32%) show loss of details and nuance, especially to omission. Serious errors (20%), although fewer, are critical because they distort meaning, remove key concepts, and reduce subtitle reliability. Overall, while the AI generated subtitles are generally readable and largely understandable, the presence of serious and standard errors demonstrate that it is not fully reliable and accurate without human revision, reinforcing prior research that emphasises the need for post-editing in machine translation output.

Readability errors such as sentence fragments, spotting, and improper line lengths were very rare. Furthermore, the model consistently avoided unnecessary sentence splitting or merging, indicating that ChatGPT followed the commands (prompts) reasonably well.

Compared with human-translated subtitles, the automatically generated subtitles do not achieve the same level of accuracy. They contain a range of errors with varying degrees of severity, including omissions, semantic and stylistic issues, mistranslations, and excessive compression. The superior quality of human-produced subtitles may stem from two main factors: the technical constraints imposed by subtitling nature, and the complexity and richness of Arabic morphology and vocabulary.

This research does not come without limitations: This preliminary study drew on presentations from only three TED TALKs, which may not fully represent the diversity of audio-visual content. In addition, the auto-generated subtitling was evaluated by a relatively small group of annotators, potentially limiting the generalisability of the findings to broader populations. The experiment also focused exclusively on a text-based chatbot - ChatGPT 3.5.

Extended research, while this study examined a standard, open-resource ChatGPT, future research could explore the performance of more advanced and multimodal ChatGPT 4/Turbo or 4o (omin) that offer larger capacity and web-browsing capabilities. Such work might incorporate a wider range of variables, including larger sample sizes, more participants,

and domain-specific analysis. Additionally, integrating automatic evaluation metrics alongside human assessment would be valuable for examining the relationship between computational measures and human judgement.

Conclusion

This study evaluated the automatically generated Arabic subtitling, with particular attention to linguistic quality and functional equivalence, using TED content. The findings reveal several issues associated with ChatGPT 3.5 generated subtitling. Human annotation identified omission as the most frequent error in AI-generated output, followed by semantic and stylistic inaccuracies, sentence compression, and mistranslation. In contrast, errors related to grammar, idiomaticity, punctuation, and spelling or redundancy were relatively rare. These issues, however, still impacted the clarity of the intended audio message and the transmission of information.

Overall, the results suggest that errors are primarily meaning- and style-related, rather than stemming from accessibility or technical limitations. The findings cannot be broadly generalized, as they may be influenced by the morphological complexity and lexical characteristics of Arabic, as well as technical constraints inherent in subtitling. Further research is needed to examine automatically generated subtitles produced by more advanced ChatGPT models.

References

Abdelaal, N., & Al Sawi, I. (2025). A comparative evaluation of machine translation vs. human translation for legal texts: a case study of translations between English to Arabic. In *Comparative Legilinguistics*, *63*, 186-223.

Al Sammarraie, H. M. (2025). Challenges and strategies in post-editing English into Arabic Neural Machine Translations of movie subtitles. In *Journal of Modern Languages*, *35*(1), 139-164.

Al Sawi, I., & Allam, R. (2024). Exploring challenges in audiovisual translation: A comparative analysis of human-and AI-generated Arabic subtitles in Birdman. In *Plos one*, *19*(10), e0311020.

Al-Darabee, M., Farghal, M., & Haider, A. S. (2025). Netflix versus Google Translate: A case study of the English-Arabic translation of scatological terms. In *Studies in Linguistics, Culture & FLT*, *13*(2), 77-95.

Al-Harahsheh, A. M., & Rababah, R. A. H. (2025). YouTube's Automated Subtitling from English into Arabic: A Case Study of Harry Potter and the Prisoner of Azkaban. In *FORUM FOR LINGUISTIC STUDIES*, 7 (2).

Al-Momani, A., Haider, A. S., Dagamseh, M., & Akasheh, W. M. (2025). Audience responses to cultural and linguistic gaps in English–Arabic auto-subtitles on YouTube. In *Research Journal in Advanced Humanities, 6*(3).

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *In International Conference on Learning Representations, (ICLR) 2015*, San Diego, California.

Bellés-Calvera, L., & Caro Quintana, R. (2022). Is academic discourse accurate when supported by machine translation. In *Quaderns de Filologia*, (171-201).

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., ... & Zampieri, M. (2016, August). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers* (pp. 131-198) Berlin, Germany.

Bywood, L., Georgakopoulou, P., & Etchegoyhen, T. (2017). Embracing the threat: machine translation as a solution for subtitling. In *Perspectives*, *25*(3), 492-508.

Calvo-Ferrer, J. R. (2024). Can you tell the difference? A study of human vs machine-translated subtitles. In *Perspectives,* 32(6), 1115-1132.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art? In *The Prague Bulletin of Mathematical Linguistics,* 108:109–120.

Cintas, J. D. (2003). Audiovisual translation in the third millennium. *Translation today: Trends and perspectives*, 192-204.

Cintas, J. D., & Remael, A. (2021). *Subtitling: Concepts and practices*. Routledge.

Farghal, T. M., & Ammari, R. M. G. (2025) Evaluating English-Arabic Automated Subtitling on Social Media Reels and Shorts: Linguistic Challenges and Performance Metrics, *Dragoman*, (21), 688-722.

Hagström, H., & Pedersen, J. (2022). Subtitles in the 2020s: The influence of machine translation. In *Journal of Audiovisual Translation*, *5*(1), 207-225.

Hiraoka, Y., & Yamada, M. (2019, August). Pre-editing plus neural machine translation for subtitling: effective pre-editing rules for subtitling of TED Talks. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, 64-72.

Hu, K., O'Brien, S., & Kenny, D. (2020). A reception study of machine translated subtitles for MOOCs. In *Perspectives*, *28*(4), 521-538.

Jiao, W., Wang, W., Huang, J. T., Wang, X., & Tu, Z. (2023). Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745*, *1*(10).

Karakanta, A., Bentivogli, L., Cettolo, M., Negri, M., & Turchi, M. (2022). Post-editing in automatic subtitling: A subtitlers' perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 261-270.

Koponen, M., Sulubacak, U., Vitikainen, K., & Tiedemann, J. (2020). MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*,115-124.

Marr, B. (2023). A short history of chatGPT: how we got to where we are today. In *Forbes*.

Martínez, J. M., & Vela, M. (2016, May). SubCo: A learner translation corpus of human and machine subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2246-2254.

Matusov, E., Wilken, P., & Georgakopoulou, Y. (2019). Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (1)*, 82-93.

Moneus, A. M., & Sahari, Y. (2024). Artificial intelligence and human translation: A contrastive study based on legal texts. *Heliyon*, *10*(6).

Moorkens, J., Way, A., & Lankford, S. (2024). *Automating translation*. Routledge.

Pedersen, J. (2017). The FAR model: assessing quality in interlingual subtitling. In *The Journal of Specialised Translation*, (28), 210-229.

Sennrich, R., Haddow, B., & Birch, A. (2016). Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: (2),* 371-376, Berlin, Germany.

Slator. (2024). Slator pro guide: Audiovisual translation. Slator Pro Guide: Audiovisual Translation - Slator

Xie, B. (2023). Machine Translation Implementation in Automatic Subtitling from a Subtitlers' Perspective. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, 54-64.

Yang, Z., Liu, F., Yu, Z., Keung, J. W., Li, J., Liu, S., ... & Li, G. (2024). Exploring and unleashing the power of large language models in automated code translation. In *Proceedings of the ACM on Software Engineering*, *1*(FSE), 1585-1608.