Google Translate in Massive Open Online Courses: Evaluation of Accuracy and Comprehensibility among Female Users

Dalal K. Alshammari
Department of English Language and Translation
College of Languages and Translation, King Saud University
Riyadh, Saudi Arabia
Corresponding Author: Dalal.Kareem.Alshammari@Gmail.com

Nasrin S. Altuwairesh
Department of English Language and Translation
College of Languages and Translation, King Saud University
Riyadh, Saudi Arabia

Received: 05/09/2022   Accepted:07/31/2022   Published: 08/24/2022

Abstract
The increasing production of audiovisual texts online has led to the growing use of Machine Translation (MT) in its raw form to facilitate multilingual access. In this study, we examine the accuracy of Google Translate’s English–Arabic subtitles and their comprehensibility among female users through a case analysis of Massive Open Online Courses (MOOCs). We also seek to contribute to the research on MT by providing empirical evidence on the use of MT in audiovisual contexts, which is a relatively new and under-researched study area. The research questions considered in this study are: (i) What types of accuracy errors can be identified in Google Translate’s English–Arabic translation of MOOC subtitles? (ii) What effects do machine-translated MOOC subtitles have on female users’ comprehension? The study used a mixed-methods approach. First, 147 machine-translated subtitles of five MOOC videos were annotated by six evaluators based on five accuracy error types: mistranslation, terminology, omission, addition, and untranslated-segment errors. Second, the comprehensibility of the machine-translated subtitles was examined through a quasi-experiment. Sixty-two female participants were divided into two groups. Each group was presented with MOOC videos with either machine- or human-translated subtitles. Then, the participants completed a comprehension test. The qualitative analysis of the accuracy errors showed the occurrence of all five accuracy error types. The quantitative analysis of users’ comprehension revealed a lack of statistically significant differences between the examined groups. The study’s results suggested that MT can be useful in facilitating users’ access to MOOC video content.

Keywords: Accuracy, audiovisual translation, comprehensibility, content understanding, evaluation, Google Translate, machine translation, manual error annotation, massive open online courses, subtitles

DOI: http://dx.doi.org/10.24093/awejtls/vol6no3.4
Introduction

Language barriers can present a significant challenge for internet users worldwide since more than half of the content online is in English (Baur et al., 2020). They are especially relevant to Arabic speakers since today’s internet is enormously lacking in Arabic content (Abubaker et al., 2015). These barriers can prevent users from accessing online written content and video materials. Many online platforms offer Machine Translation (MT) to eliminate these obstacles. MT in Massive Open Online Courses (henceforth MOOCs) is a prime example of how MT can enable multilingual access in written and video forms.

The use of MT to translate video subtitles is an example of multimodal MT, which involves translating sentences situated in a visual context (Elliott, 2018). It is an interesting area of study, especially in terms of accuracy and comprehensibility. One reason for this is that, unlike written MT, the quality of multimodal MT depends not only on the linguistic elements in the Source Text (ST) but also on the semiotic elements in the image or scene. The different semiotic layers in audiovisual material (e.g., images, gestures, music, and noise) can add more complexity to the translation task.

The present study seeks to contribute to the research on MT by providing empirical evidence on the use of MT in audiovisual contexts, which is a relatively new and under-researched study area. Although the quality of MT has always been a concern for researchers, many studies conducted on this subject focus on written texts (Klubička, Toral, & Sánchez-Cartagena, 2018; Stasimioti & Sosoni, 2019; Carl & Báez, 2019; Castilho et al., 2017; Guerberof-Arenas & Toral, 2022). Thus, we aim to fill this gap by conducting a qualitative analysis of MT when used for the translation of audiovisual texts. Additionally, we provide a better understanding of MT by examining female users’ comprehension of the produced output, as research on users’ comprehension of MT outputs is significantly limited (Koehn, 2020).

The purpose of this study is twofold: (i) to evaluate the accuracy of Google Translate when used for the translation of MOOC subtitles through error annotation and (ii) to examine the comprehensibility of the selected machine-translated MOOC subtitles among female users through a quasi-experiment. Using a mixed methods approach, this study addresses the following research questions:

RQ1: What types of accuracy errors can be identified in Google Translate’s English–Arabic translation of MOOC subtitles?
RQ2: What effects do machine-translated MOOC subtitles have on female users’ comprehension?

This paper is structured as follows: first, we review the literature on MT in MOOCs, MT evaluation, and Google Translate’s efficiency with Arabic texts. Afterward, we outline the
methodology by presenting the method, participants, research instruments, and research procedures. We then present the results, followed by a discussion of the major themes in the qualitative and quantitative analyses. Finally, we present the conclusions drawn from this paper.

Literature Review

**Machine Translation in Massive Open Online Courses**

E-learning platforms, including MOOCs, have gained considerable attention worldwide due to their increased availability and access. The term “Massive Open Online Course” was coined in 2008 to refer to “an online course aimed at unlimited participation and open access via the web” (Faizuddin & Azeeza, 2015, p. 1). Many elite universities and organizations offer MOOCs to anyone willing to enroll. However, several recognized MOOC platforms adopt English as their language of instruction, which could be challenging for learners who speak foreign languages.

To overcome this issue and enhance learners’ access to MOOC content, many initiatives have been introduced to provide MT services for online courses, including the European Commission-funded TraMOOC project, the European Multiple MOOC Aggregator (EMMA) project, and the Qatar Computing Research Institute Educational Domain Corpus.

The aim of the TraMOOC project was to make MOOCs more accessible by developing a high-quality translation tool (Bethge et al., 2021). The project established a neural MT (NMT) platform to translate MOOCs into 11 languages: Bulgarian, Chinese, Croatian, Czech, Dutch, German, Greek, Italian, Polish, Portuguese, and Russian (Castilho, Gaspari, Moorkens, & Way, 2017). According to Bethge et al. (2021), the translation quality provided by this project was generally perceived well by course participants and instructors.

Another effort to broaden the accessibility of MOOC content is the EMMA project. The project provided a system for delivering free MOOCs from various European universities with integrated automatic translation systems (Miró, Baquero-Arnal, Civera, Turró, & Juan, 2018). Its website provides automatic transcription in seven languages: Dutch, English, Estonian, French, Italian, Portuguese and Spanish. It also offers MT in three languages: English, Italian, and Spanish (Hu, 2020). Lecturers and learners review the transcriptions and translations to ensure that they are of publishable quality (Miró et al., 2018).

Similarly, the Qatar Computing Research Institute made online educational content accessible in over 20 languages, including Arabic, by introducing the Educational Domain Corpus. According to Abdelali, Guzman, Sajjad, and Vogel (2014), this corpus is an open multilingual collection of subtitles for educational videos and lectures collaboratively transcribed and translated over the AMARA web-based platform. It was built using statistical MT (SMT) and manual translations.
In addition to the initiatives mentioned above, many well-known platforms, such as Udacity and Khan Academy, adopt Google Translate to enable learners to view content in their native languages. Google Translate is a free online service and a neural-based translation tool; thus, it relies on predictive algorithms that use complex neural networks to produce the translated output (Wu et al., 2016). Although Google Translate can serve as a practical translation tool in many MOOC platforms, many questions are still raised about its quality. The issue of quality is especially pertinent to Arabic and English due to the distinct differences in their semantic and syntactic structures. Hence, evaluating Google Translate’s English–Arabic MOOC subtitles is paramount.

**Machine Translation Evaluation**

Translation quality assessment is one of the key issues in translation studies. As stated by Saldanha and O’Brien (2014), “it allows us to measure the impact and effect of different variables on the translation product and process and to subsequently change our techniques, training, or tools in order to better meet quality requirements” (p. 96). However, the concept of quality is difficult to measure or define precisely. As a result, scholars have adopted varying approaches to describe and investigate the quality of translation.

MTs are usually evaluated based on three main approaches: manual evaluation, automatic evaluation, and task-based evaluation (Koehn, 2020). Generally, MT evaluation is often performed within the borders of computer science (Hu, 2020). In other words, evaluation tests and metrics are conducted by specialists examining the performance of MT engines. However, more connections have recently been made between MT research and translation studies. These connections have been established by examining some aspects of MT related to translation research, such as the impact translation tools have on the translated product, translation process, and translators' workflow (O’Brien, 2013).

**Automatic Machine Translation Evaluation**

Automatic evaluation metrics (AEMs) use “an explicit and formalised set of linguistic criteria to evaluate the MT output, typically against a human reference translation” (Doherty, 2017, p. 4). Although there are numerous AEMs, the most notable are word error rate (WER), translation error rate (TER), and the bilingual evaluation understudy (BLEU) metrics.

The WER metric is a simple automatic metric commonly used in automatic speech recognition (Díaz-Munío, 2020). It was one of the first AEMs used in MT evaluation (Castilho et al., 2018). Despite its simplicity, the WER is fundamental to understanding several AEMs (Díaz-Munío, 2020). In simple terms, WER expresses the ratio of the required word-level editing operations for the MT output to match the human translation reference (Castilho et al., 2018). It computes the number of insertions, deletions, and substitutions divided by the length of the human
translation reference, which results in a score between zero and one, represented as a percentage (Díaz-Munío, 2020).

Another well-known metric is the TER introduced by Global Autonomous Language Exploitation (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006). Similar to WER, TER expresses the ratio of the required word-level editing operations used for the MT output to match the human translation reference. It accounts for the same errors as the WER metric: the number of insertions, deletions, and substitutions, in addition to word order. In other words, the number of required editing operations is divided by the length of the human translation reference translation. It often results in a score between zero and one, expressed as a percentage (Díaz-Munío, 2020).

In addition to the previously mentioned metrics, the BLEU metric is also one of the most recognized AEMs (Papineni et al., 2002). As with most AEMs, this method determines translation quality by computing the MT output’s closeness to the human translation reference. It measures the degree of overlap between the machine-translated output and the human reference translation by comparing isolated words and sequences of words. In essence, BLEU computes an n-gram precision, that is, the percentage of n-grams from the MT that match the human translation reference (Al-Kabi, Hailat, Shawakfa, & Alsmadi, 2013). An n-gram is a sequence of n items; in the case of MT, it refers to word sequences. The BLEU metric provides a score from zero to one, expressed as a percentage. Scores closer to one indicate a high n-gram precision (Al-Kabi et al., 2013).

AEMs are widely utilized in research for their simplicity and practicality. However, in a critique of these metrics, Doherty (2017) observed that because they mostly focus on a narrow range of formal linguistic features, they are not sophisticated enough to address the issues of fluency and accuracy. Marín Buj (2017) also suggested that automatic metrics often require additional human contribution.

Manual Machine Translation Evaluation

In manual evaluation metrics, the quality of the output is judged by a human evaluator. Most manual metrics are based on two main categories: fluency and accuracy, with the latter also known as adequacy (Lommel, Uszkoreit, & Burchardt, 2014; Maučec & Donaj, 2019). Koehn (2020) defined fluency as the syntactic correctness of the MT output in terms of grammatical structures and idiomatic word choices. He also explained that accuracy refers to conveying the semantic meaning of the ST and deals explicitly with loss, gain, or distortion in meaning.

Many manual evaluation taxonomies have been proposed. These taxonomies aim to provide quantifiable evaluation criteria. They often involve counting, classifying, and weighing errors according to their severity (Castilho et al., 2018).
The Localization Industry Standards Association (LISA)’s quality assurance (QA) model, developed in 1995, was one of the first manual error-based metrics (Martínez, 2014). It sought to optimize translation and localization methods for the software and hardware industries (Görög, 2017). The first LISA QA model was based on pre-existing quality metrics used by many companies, such as Microsoft and Oracle (Sprung & Jaroniec, 2000). Since 2011, the LISA QA has no longer been active, although the organization’s standardization methods are still widely used in translation quality evaluation (Görög, 2017).

The dynamic quality framework (DQF) is another notable error-based metric. It was developed in 2011 by the Translation Automation User Society. It was designed primarily based on pre-existing metrics, with the LISA QA playing a central role in its development (Görög, 2014). However, unlike the static checklist of the LISA QA and previous error-based metrics, the DQF is implemented in a dashboard, which serves as a translation quality and assessment knowledge base (Marín Buj, 2017).

More recently, the multidimensional quality metric (MQM) was developed in 2015 as part of the European QTLaunchPad project (Lommel et al., 2014). Like the above-mentioned error-based metrics, it is considered an extension of pre-existing evaluation frameworks. However, the MQM provides a wide range of error categories that enable flexibility and customization. As asserted by Lommel et al. (2014), researchers are not expected to utilize all the categories in the metric and should instead make relevant selections according to the text intended for evaluation.

Several studies have adopted manual evaluation metrics to evaluate the quality of MT. Castilho et al. (2017) conducted a comparative evaluation of phrase-based SMT and NMT for four language pairs. They used a corpus of machine-translated MOOC material in German, Greek, Portuguese and Russian. The evaluation included a variety of metrics: automatic evaluation, human rankings of adequacy and fluency, error-type annotation, and post-editing. Human evaluation was performed by professional translators. For the error-type annotation task, the annotators were asked to annotate the errors based on a simple error taxonomy comprising inflectional morphology, word order, omission, addition, and mistranslation errors. The results of the ranking task showed moderate inter-annotator agreement among the annotators and a preference for NMT for all language pairs. NMT was also perceived to have better fluency and fewer annotated errors across all languages. However, the results were mixed for perceived adequacy and omission, addition, and mistranslation errors. Inflectional morphology, word order, and mistranslation were the most frequent error types in both MT systems.

Klubička et al. (2018) conducted a case study to compare the output of three MT systems: pure phrase-based, factored phrase-based, and NMT systems. They examined the translation of 100 randomly selected sentences from various publicly available corpora of English–Croatian
translations. The data were presented to two annotators who were asked to annotate the errors according to a tailored MQM error taxonomy. The researchers concluded that the NMT system contained the smallest number of errors. They found that the most common error type in the NMT output was mistranslation, followed by grammar and word form errors. They also noted a striking drop in agreement scores for the NMT system. The researchers attributed the low agreement to the increased fluency of the NMT, which made error detection challenging for the annotators.

Similarly, Stasimioti and Sosoni (2019) compared the performance of different MT systems using various evaluation methods. They examined four short, semi-specialized texts about the 2019 EU elections selected from the Guardian newspaper. The articles were translated from English into Greek using three MT systems: Google Translate SMT system, Google Translate NMT system, and a tailored-NMT system. The comparison was performed following a mixed-methods approach consisting of automatic evaluation metrics, manual ranking, error classification, and post-editing effort. The error classification task was based on a combination of the subsets of the DQF and MQM error taxonomies. The findings revealed that both the generic NMT and the tailored NMT outputs scored higher on the human evaluation metrics. As for the error-type classification, grammar errors were the most prevalent in all MT systems, followed by mistranslation and style errors. The inter-annotator agreement showed fair agreement among the annotators for fluency and slight agreement for accuracy.

In addition, Carl and Báez (2019) investigated the manual error annotation and post-editing effort of English–Spanish and English–simplified Chinese MT output generated by Google Translate. The STs were collected from previous studies based on Google Translate’s output. Sixteen Chinese translation students and eight professional Spanish translators were asked to annotate the outputs of the same English STs based on an MQM-derived error taxonomy. Each annotated error was marked as either critical or minor. The findings suggested that the English–Spanish MT contained more accuracy errors and fewer fluency errors than the English–simplified Chinese MT.

Guerberof-Arenas and Toral (2022) also used manual evaluation methods to examine creativity in three modalities: MT, post-edited MT, and human translation. The collected data included the translations of a short story by Kurt Vonnegut from English to Catalan and Dutch. The analysis relied on two criteria: novelty and acceptability. To measure the translations’ acceptability, the researchers used the DQF–MQM error typology. The results revealed that human translation had the highest creativity score, followed by post-edited MT and MT. For the error analysis, the researchers concluded that MTs are often too literal and rife with mistranslations, causing them to become impractical.
Task-based Machine Translation Evaluation

Given that MT is utilized by a large audience of diverse languages, backgrounds, and age groups, it is imperative to address the receivers’ role in judging the quality of the output. Information on how users perceive the quality of the output can be obtained through various empirical task-based evaluation metrics. What distinguishes task-based metrics from automatic and manual methods is that they often go a step further from linguistic analysis to measure the extent to which the MT output fulfills its intended purpose in real-life settings (Koehn, 2020).

Free online MT engines are widely used to serve a variety of purposes. Hutchins (2007) briefly summarised the different applications of MT as assimilation (to roughly understand the text in another language), dissemination (to publish a text in several languages for a wide audience), and communication (to facilitate interactions between people who speak different languages). Overall, MT is not expected to provide a well-translated output. Instead, it is often used to produce an instant translation that conveys the message of the original text, however imperfect or linguistically awkward (Hutchins, 2007).

Koehn (2020) suggested two main task-based metrics: translator productivity and content understanding. Translator productivity is often measured by the effort required for post-editing, whereas content understanding is measured by how well the user of an MT system understands the translated text.

MT can be used to increase the productivity of professional translators by producing the first draft for post-editing and publication. Post-editing is the correction of raw machine-translated output by a human translator according to specific guidelines and quality criteria (O’Brien, 2011). To measure the success of MT in increasing translator productivity, researchers often examine how much time professional translators spend on post-editing the MT output. Generally, translator productivity metrics seek to determine whether specific MT systems can save time, effort, and cost by providing an instantaneous translation that requires few corrections.

Content understanding metrics are used to measure users’ understanding of the essential information in the MT output. If the MT output permits a complete understanding of the presented information, it is considered to be of high quality. If it only enables an understanding of the general topic, it is considered to be of poor quality. Overall, the aim of content understanding metrics is to examine whether a monolingual target language speaker can answer questions about a presented MT output (Koehn, 2020). Although task-based MT evaluation is not recent, research on content understanding is limited. These evaluation metrics are especially worthy of further exploration, considering that there have not been many large-scale studies in which they are addressed (Koehn, 2020).
Scarton and Specia (2016) explored users’ understanding of MT. The researchers’ main aim was to create a corpus that included translated German reading comprehension texts and exercises. To evaluate the corpus, they conducted a study to assess the scores of participants who were presented with German machine-translated documents. They examined the scores obtained from 19 sets of either machine-translated or human-translated documents. Each document was given to each participant. The comprehension questionnaire included both open- and closed-ended questions. The researchers concluded that the scores varied among the test-takers, and the agreement rate was low; therefore, they stated that drawing significant conclusions was difficult.

Forcada, Scarton, Specia, Haddow, and Birch (2018) used the same corpus discussed in Scarton and Specia’s (2016) study to examine users’ understanding of MT. In Forcada et al.’s (2018) work, 30 participants were given a set of six documents, either machine or human translated, and asked to answer three to five questions per document in their native language. The documents were translated using four MT systems: Google Translate, Bing, Homebrew, and Systran. The reading comprehension questions included both objective and inference questions. The results showed that the users' scores obtained with professional human- and machine-translated documents were not statistically different. Among the MT systems, Google Translate had the highest user scores. However, as the research examined MT from English to German, the results may differ for other language pairs.

Castilho and Guerberof-Arenas (2018) carried out a pilot experiment to compare users’ reading comprehension of SMT and NMT in different languages. To collect the data, they first presented three machine-translated IELTS General Training reading texts to six native speakers of English, Spanish and simplified Chinese. Afterward, they asked the users to answer the respective comprehension questions and rank their satisfaction with each text on a three-point scale. The researchers also used an eye tracker to gather quantitative data on the users’ reading habits and conducted post-task interviews. The findings suggest that the users completed more tasks in less time, with a higher level of satisfaction when using the NMT system. However, given the small number of participants, and as stated by the researchers, the study focused on verifying the used methods rather than obtaining significant results.

Hu (2020) undertook a similar experiment by examining the comprehension of machine-translated subtitles on MOOC platforms. She used a mixed-methods approach consisting of eye tracking, translation quality assessment, and questionnaires. The study included 60 participants who were native speakers of Chinese. The participants were divided into three groups. Each group was presented with either human-translated, machine-translated, or post-edited machine-translated subtitles. The researcher used a comprehension questionnaire that included multiple-choice and true/false questions. The results of the comprehension questionnaire showed that the performance
of the participants who were given post-edited MT and human-translated subtitles was better than that of the groups that were given machine-translated subtitles.

Fantinuoli and Prandi (2021) adopted a different methodology to examine MT comprehensibility. They compared the English–Italian MT and the transcription of the translations produced by human interpreters based on two dimensions: intelligibility (i.e., target-text comprehensibility) and informativeness. Six participants with a background in interpreting and translation were asked to rate the translations according to these dimensions on a six-point Likert scale. The researchers observed that human interpreters often intervene in the translation to improve the comprehensibility of the text. In general, the study revealed better performance by human interpreters in terms of intelligibility and slightly better performance by the MT engine in terms of informativeness. However, a clear limitation of this study is its reliance on text analysis rather than on the reception of actual MT users.

Google Translate’s Efficiency with Arabic Texts

Google Translate was introduced in 2006 and is considered one of the most popular machine translation systems (Sabtan, Hussein, Ethelb, & Omar, 2021). It has become widely used worldwide, serving approximately 200 million users daily (Habeeb, 2020). It may be preferred in many cases where there is a need for inexpensive and fast translations. However, translations between distinct languages, such as English and Arabic, can be challenging for machines. Therefore, assessing Google Translate’s output when dealing with Arabic texts is crucial.

There have been several small-scale studies on Google Translate’s English–Arabic translation. Diab (2021) compared the quality of Google Translate’s Arabic NMT and SMT using manual evaluation methods to identify and classify errors generated by each MT model. The analysis was guided by the DQF–MQM error typology, and the data used in the study comprised a corpus of six English WikiHow articles on cybersecurity, cryptocurrency, and healthcare. The results showed that the NMT engine reduced the number of errors in the Arabic output by almost 80% compared to the SMT. Thus, Google Translate’s NMT can be regarded as being more efficient and reliable with Arabic texts.

In terms of NMT systems, Abu-Ayyash (2017) compared the Arabic MT outputs of three NMT systems: Systran’s Pure NMT, Google Translate, and Microsoft Bing Translator. The analysis included the evaluation of errors in the translation of gender-bound constructs (i.e., subject-verb agreement, adjectival-noun agreement, and pronoun-antecedent agreement) in four selected technical texts that were machine-translated from English to Arabic. The study indicated that Systran’s Pure NMT performed better than the other two MT systems. However, the researcher concluded that these results could not be generalized due to the limited sample.
Similar to Abu-Ayyash’s (2017) study, Al-Jarf (2021) examined the accuracy of Google Translate's Arabic translation of English technical terms and discussed the semantic, contextual, syntactic, morphological, and orthographic deficiencies in the output. The study included a sample of technical terms with different Greek and Latin roots and affixes. The results suggested that Google Translate provided Arabic equivalents to some terms but is inconsistent in translating terms with varying affixes, compounds, and blends.

Furthermore, Almahasees (2017) addressed the adequacy and fluency of Google Translate's output compared to Microsoft Bing Translator using the BLEU AEM. The researcher examined the English translation of Khalil Gibran's Arabic literary work, "The Prophet," and the analysis showed that the MTs of both systems were inaccurate in translating literary texts due to the difficulty of literary language. Literary texts often include metaphors, cultural specifications, sophisticated lexical terms, and syntactic structures that can be difficult to render. The author also found that both systems provided similar translations for the same input.

Moreover, Abdelaal and Alazzawie (2020) analyzed the errors in Google Translate’s English output of Arabic informative news texts. The aim of the research was to measure the translation quality to determine the extent to which human post-editing is needed. The error analysis was based on the MQM metric, in addition to the following error categories: orthographic, morphological, lexical, semantic, and syntactic errors. The study revealed that omissions and inappropriate lexical choices were the most common. The researchers concluded that although MT could help expedite the translation process, the accuracy of the translated text could be compromised.

Jabak (2019) also examined Google Translate’s Arabic–English output by assessing a sample of translations from a book entitled "Thinking Arabic Translation" against model translations. The researcher developed a qualitative error analysis method to categorize the identified errors into different themes based on their nature and recurrence. The findings indicated that Google Translate made lexical and syntactic errors, which affected the quality of the translated text and made it incomprehensible.

Methods

One of the major criticisms of earlier audiovisual translation research is its limited and partial focus on linguistic and cultural matters, despite the multifaceted nature of audiovisual translation (Gambier, 2009). From this standpoint, this research integrates various theoretical perspectives, specifically manual and task-based evaluation metrics, to provide a comprehensive insight into Google Translate’s translation of subtitles through case study analysis. More specifically, this research adopts a mixed-methods approach, utilizing qualitative content analysis to determine accuracy errors in Google Translate's machine-translated subtitles and statistical tests.
to compare differences in female users' comprehension of machine- versus human-translated subtitles.

**Participants**

The main weakness of manual evaluation metrics is their high level of subjectivity (Koehn, 2020). Although researchers in the field of MT have attempted to define the evaluation criteria and categories of manual metrics, they are often difficult to put into practice because clear-cut boundaries may not be apparent between each category. To overcome this issue in the present study, translation experts examined the machine-translated output of each video twice. The participants consisted of Saudi female postgraduate students in translation studies (n = 6). They were chosen through convenience sampling for their expertise in and knowledge of translation, including translation quality assessment.

For the quantitative experiment on content understanding, the sample consisted of first-year Saudi female university students (n = 62). The participants' ages ranged from 19 years to 23 years, and they were all native speakers of Arabic. Initially, the participants were selected randomly through convenience sampling to reach a more diverse population. However, considering that the command of foreign languages can affect the reception of translated subtitles (Gambier, 2009), the sample was further stratified through an English proficiency test and distributed equally into an experimental and a control group according to the participants’ proficiency levels.

**Research Instruments**

Many recognized MOOC platforms offer machine-translated subtitles to facilitate multilingual access. In the present work, MOOCs were chosen for the case study because of their growing popularity and increased availability. Furthermore, they are often informative in nature; thus, investigating the accuracy of the translated information, especially in relation to users' comprehension, may be of particular importance. The course selected for this study consisted of eight short videos (i.e., 1–3 min each). It focuses on a general skill-based topic. The course is available for free on Udacity and was produced by the Grow with Google initiative (Udacity, n.d.). As the participants were expected to watch the videos and complete a comprehension test in one sitting, the first three introductory videos were not included, and only five videos were used in this research. The selected videos were 12:15 min in total. The machine-translated subtitles of the videos were produced by Google Translate. The subtitles were used for both the qualitative annotation of accuracy errors and the quantitative quasi-experiment on content understanding. Additionally, the researchers provided professional human-translated subtitles for the quasi-experiment.
For manual error annotation, five evaluation forms were created as Microsoft Word documents to facilitate the evaluators’ tasks. The forms contained a description of the accuracy error types (Table one), in addition to a table showing the original English subtitles and their corresponding MTs. Each form included subtitles for one video.

Table 1. *Definitions of the accuracy error types in the MQM model*

<table>
<thead>
<tr>
<th>Accuracy Error</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mistranslation</td>
<td>The TT content does not accurately represent the ST content.</td>
</tr>
<tr>
<td>Terminology</td>
<td>Terminology issues relate to the use of domain- or organization-specific terminology.</td>
</tr>
<tr>
<td>Omission</td>
<td>Content that is present in the ST is missing from the TT.</td>
</tr>
<tr>
<td>Addition</td>
<td>The TT includes text not present in the ST.</td>
</tr>
<tr>
<td>Untranslated-segment</td>
<td>Inaccurate direct transference of segments from the ST into the TT without any change in their form is present.</td>
</tr>
</tbody>
</table>

*Note 1. Adapted from Hu (2020, p. 333)*

The comprehension test for the quasi-experiment was designed to cover both factual and inference questions and consisted of 15 items. These items were related to the content of the presented videos and covered a range of difficulty levels. More specifically, the test included true-or-false and multiple-choice questions. The questionnaire was initially developed in English by the researchers to ensure that the questions matched the presented information in its original language. Considering that most participants had low to intermediate levels of English proficiency, the test was translated by the researchers into Arabic and presented to the participants in their native language. Google Forms was used to present both the videos and the comprehension test.

*Research Procedures*

Data were collected during the spring 2021 semester. For the evaluation task, the evaluators (n = 6) were given a form detailing the evaluation criteria and the original English subtitles with their corresponding MTs. The error annotation process was explained in detail, and the participants’ questions were addressed through discussion. The videos selected for this research ranged in length from one to two minutes. Each participant was given subtitles to two short videos (i.e., 1 min each) or one longer video (i.e., 2 min), as shown in Table two. Error annotation was performed segment by segment using Microsoft Word. The evaluation forms were intentionally distributed so that each evaluator would not be with the same person twice to ensure the reliability of the evaluations (Mariana, Cox, & Melby, 2014). As the machine-translated subtitles of each video were presented to two evaluators, the interrater agreement percentage was also measured for further analysis. The agreement percentage was measured as the number of times the evaluators agreed on an error type divided by the total number of errors and then multiplied by 100.
For the comprehension test, the participants (n = 62) were randomly selected through convenience sampling. The sample was then stratified using a shorter version of the Cambridge General English Test (i.e., consisting of 25 questions) to eliminate any significant differences in English proficiency between the participants of the control and experimental groups. The participants were met online via Zoom. Those in the control group were given a form that included the selected MOOC videos with human-translated subtitles, whereas those in the experimental group were given another form that included the same videos with machine-translated subtitles. The participants in both groups completed the same comprehension test.

The data collected through the comprehension test were then analyzed statistically using SPSS version 28.0. A Kolmogorov–Smirnov normality test (Table three) indicated that the scores of the experimental group were not normally distributed. Thus, an independent samples $t$-test was considered inapplicable. Alternatively, a Mann–Whitney $U$ test was implemented in this study to compare the mean ranks rather than medians.

Table 3. Kolmogorov–Smirnov test of normality

<table>
<thead>
<tr>
<th>Group</th>
<th>Statistic</th>
<th>N</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>.134</td>
<td>31</td>
<td>.163</td>
</tr>
<tr>
<td>Experimental</td>
<td>.227</td>
<td>31</td>
<td>.000</td>
</tr>
</tbody>
</table>

Results

Qualitative Analysis

The qualitative analysis showed 131 accuracy errors in the 147 subtitle segments analyzed. The most prevalent error type was mistranslation, which accounted for 74% of all translation errors, as depicted in Figure one. Terminology errors were the second most common, although they accounted for only 14% of the errors. Omission and addition errors were less frequent; each comprised 5% of all errors. Untranslated-segment errors occurred in only two instances, accounting for 2% of the identified errors.

Figure 1. Percentages of identified accuracy errors
An error annotation of each video subtitle was carried out by more than one evaluator to limit the subjectivity of manual evaluation by offering multiple assessments of the same texts. However, considering that manual evaluations are highly subjective (Klubička et al., 2018), an additional analysis was carried out by the researchers to measure interrater agreement among the evaluators. The data presented in Table four reveal that the two instances of untranslated-segment errors had the highest agreement percentage (100%) among the evaluators. Terminology errors had the second-highest agreement percentage (42%). Mistranslation and addition errors, however, had low agreement percentages of 26% and 14%, respectively. The evaluators did not agree on any omission errors. Overall, the agreement percentage across all error types was approximately 27%.

Table 4. Frequencies and percentages of agreement on identified accuracy errors

<table>
<thead>
<tr>
<th>Accuracy Error Type</th>
<th>Error Frequency</th>
<th>Percentage</th>
<th>Agreement Frequency</th>
<th>Agreement Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mistranslation</td>
<td>97</td>
<td>74.04</td>
<td>25</td>
<td>25.77</td>
</tr>
<tr>
<td>Terminology</td>
<td>19</td>
<td>14.50</td>
<td>8</td>
<td>42.10</td>
</tr>
<tr>
<td>Omission</td>
<td>6</td>
<td>4.58</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Addition</td>
<td>7</td>
<td>5.34</td>
<td>1</td>
<td>14.28</td>
</tr>
<tr>
<td>Untranslated-segment</td>
<td>2</td>
<td>1.52</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>131</td>
<td>100</td>
<td>36</td>
<td>27.48</td>
</tr>
</tbody>
</table>

Quantitative Analysis

To explore the differences in the scores of the control and experimental groups, we implemented a Mann–Whitney U test. An alpha level of 0.05 was used. The results indicated no statistically significant differences between the control and experimental groups; \( U = 471.000, z = -0.136, \) and \( p > .05 \) (Table five). The mean ranks of the control and experimental groups were 31.81 and 31.19, respectively. This suggests that the ranks of the scores were similar for both.
groups. Thus, users’ comprehension was not negatively affected by differences in the level of accuracy of human-translated subtitles and Google Translate’s machine-translated subtitles.

Table 5. Results of the Mann–Whitney U Test

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean Rank</th>
<th>Sum of Ranks</th>
<th>U</th>
<th>Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>31</td>
<td>31.81</td>
<td>986.00</td>
<td>471.000</td>
<td>-.136</td>
<td>.892</td>
</tr>
<tr>
<td>Experimental</td>
<td>31</td>
<td>31.19</td>
<td>967.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

Qualitative Analysis

The first research question of this paper is as follows:

RQ1: What types of accuracy errors can be identified in Google Translate’s English–Arabic translation of MOOC subtitles?

To answer this question, six evaluators annotated the accuracy errors in the machine-translated outputs according to the following error types in the MQM model: mistranslation, terminology, omission, addition, and untranslated-segment errors. The results of the qualitative analysis showed the occurrence of all five accuracy error types in the MQM model: mistranslation, terminology, omission, addition, and untranslated-segment errors. Mistranslation errors were the most common. This result supports Klubička et al.’s (2018) finding that mistranslation errors are the most frequent, comprising most of the identified errors in the examined outputs that are translated from English into Croatian using an NMT engine. Likewise, Castilho et al. (2017) found that mistranslation errors were more common than other types of accuracy errors in the examined corpus, which comprised NMTs in German, Greek, Portuguese and Russian. Carl and Báez (2019) also suggested that mistranslation errors are more common than other types of errors in both English–Spanish and English–simplified Chinese MT outputs generated by Google Translate.

In the current study, terminology errors were the second most common. However, similar to Hu’s (2020) findings, these errors were less frequent because the analyzed material was not specialized. Most of the identified terminology errors were related to over-literal translation, which is expected in MT outputs. Omission and addition errors infrequently occurred in the examined data, which was also expected because MTs often depended solely on linguistic input. As evident in this study, Google Translate’s engine rarely omitted any part of the text, compensated for any omissions in the ST, or added any new information. Untranslated-segment errors were the least frequent. They occurred in the translation of only two proper nouns. This finding was also expected and is in line with that of Klubička et al. (2018), which indicated the rare occurrence of untranslated-segment errors. In summary, the findings of the present study regarding the frequency of error types are consistent with those of the reviewed literature.
The overall agreement percentage among the evaluators was nearly 27%, which was expected to be low. Manual evaluation metrics are expected to be subjective. If given more time, the evaluators themselves may perceive errors differently than they had previously annotated (Klubička et al., 2018). Hu (2020) argued that, even in cases in which evaluators are trained professionally, intra- and inter-rater agreement can still be low. A review of studies that used manual evaluation showed that Stasimioti and Sosoni’s (2019) findings have slight agreement on accuracy errors. Similarly, Klubička et al. (2018) noted strikingly low agreement in the annotation of errors in NMT outputs.

Quantitative Analysis

The second research question of this paper is as follows:

RQ2: What effect do machine-translated MOOC subtitles have on female users’ comprehension?

To explore this question, we implemented a quasi-experimental comparison using a Mann–Whitney U test. The results indicated no significant differences between the two examined groups. Evidently, the machine-translated subtitles were successful in conveying some of the presented information to the participants. The findings of the quantitative analysis are consistent with those of Forcada et al. (2018), who reported a lack of any significant differences in reading comprehension between participants who were provided with different MT outputs and those who were provided with human translations. The researchers stated that this result is a clear indication of the usefulness of raw MT output in assimilating the gist of the presented information. This statement holds true in the case of this research, as the participants in the experimental group understood the MT output almost equally as well as the control group. The researchers also stated that the participants who were presented with Google Translate translations performed best among the other participants who were presented with the outputs of different MT engines, which further supported the findings of this study.

By contrast, Hu (2020) suggested that the performance of participants who were given raw MT was worse compared with that of the participants who were given post-edited MT and human-translated subtitles. Hence, we can assume that the accuracy of the machine-translated output can be lower with certain language pairs. As Callison-Burch et al. (2010) highlighted, the content understanding of MT can be affected by the output’s language. Carl and Báez (2019) also suggested that the frequency of accuracy and fluency errors in MT can vary according to the examined language pairs, which may affect users’ comprehension. As stated by Guerberof-Arenas and Toral (2022), a high number of errors can cause the translation to be impractical.

Castilho and Guerberof Arenas (2018) found that users who were given texts that were translated using NMT completed more reading comprehension tasks in less time and had a higher level of satisfaction compared with users who were given texts that were translated using SMT.
Interestingly, their study also indicated that users who were provided with MTs produced by an NMT system in Spanish and simplified Chinese scored slightly higher than native English speakers who were provided with the original STs. Scarton and Specia (2016) also noted that the scores of users who were given human-translated documents showed lower agreement compared with the scores of users who were presented with the outputs of different MT systems. These results suggest that NMT systems can produce outputs that have an acceptable level of accuracy compared with texts written in their original language. This can be the case, as Diab (2021) noted that NMT systems have significantly higher accuracy with Arabic texts compared to SMT systems.

Conclusion
This study set out to investigate the accuracy of Google Translate’s English–Arabic subtitles and the subtitles’ comprehensibility among female users through a case analysis of MOOCs. The findings indicated that Google Translate's English–Arabic translation of MOOC subtitles could be accurate and well received by users. The results of the qualitative content analysis highlighted the low number of accuracy errors in the subtitle segments examined, with a frequency of one error per subtitle segment. They also showed that most of the identified errors were related to mistranslation and stemmed from the over-literal translation of some words and phrases. In the quantitative analysis of the quasi-experimental comparison, the results indicate the lack of any significant differences in female users' comprehension of machine- and human-translated subtitles. Overall, the use of MT in audiovisual contexts is not explored in sufficient depth, especially for certain language pairs, such as English–Arabic. Thus, further investigation into the quality of machine-translated subtitles is recommended.

Acknowledgments
The authors would like to thank the Deanship of Scientific Research (DSR) at King Saud University and the College of Languages and Translation Research Center for funding and supporting this research through the initiative of DSR Graduate Students Research Support.

About the Authors
Dalal Alshammari is an English–Arabic translator and language interpreter. She has professional experience in translation and interpreting, as well as research experience in the field of translation studies. She holds a Master of Arts in Translation from King Saud University. Her research interests include audiovisual translation, machine translation, and translation quality assessment. ORCID id: https://orcid.org/0000-0002-4401-408X

Nasrin Altuwairesh is Assistant Professor at the College of Languages and Translation (COLT) at King Saud University in Riyadh- Saudi Arabia. She holds a PhD in Education/ TESOL from the University of Leeds in the United Kingdom. She has taught a number of translation courses.
at COLT. Her research interests focus on language teaching as well as the teaching of translation and training translators. ORCID id: https://orcid.org/0000-0003-0823-0743

References
Google Translate in Massive Open Online Courses

Alshammari & Altuwairesh


