International Journal of English Linguistics; Vol. 10, No. 4; 2020
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Neural Machine Translation: Fine-Grained Evaluation of Google Translate Output for English-to-Arabic Translation

Linda Alkhawaja[1], Hanan Ibrahim[1], Fida' Ghnaim[1] & Sirine Awwad[1]

[1] Department of English Language and Translation, Al-Ahliyyah Amman University, Amman, Jordan

Correspondence: Linda Alkhawaja, Department of English Language and Translation, Al-Ahliyyah Amman University, Amman, Jordan. E-mail: L.alkhawaja@ammanu.edu.jo

Received: March 10, 2020 Accepted: April 16, 2020 Online Published: April 27, 2020

doi:10.5539/ijel.v10n4p43 URL: https://doi.org/10.5539/ijel.v10n4p43

Abstract

The neural machine translation (NMT) revolution is upon us. Since 2016, an increasing number of scientific publications have examined the improvements in the quality of machine translation (MT) systems. However, much remains to be done for specific language pairs, such as Arabic and English. This raises the question whether NMT is a useful tool for translating text from English to Arabic. For this purpose, 100 English passages were obtained from different broadcasting websites and translated using NMT in Google Translate. The NMT outputs were reviewed by three professional bilingual evaluators specializing in linguistics and translation, who scored the translations based on the translation quality assessment (QA) model. First, the evaluators identified the most common errors that appeared in the translated text. Next, they evaluated adequacy and fluency of MT using a 5-point scale. Our results indicate that mistranslation is the most common type of error, followed by corruption of the overall meaning of the sentence and orthographic errors. Nevertheless, adequacy and fluency of the translated text are of acceptable quality. The results of our research can be used to improve the quality of Google NMT output.

Keywords: Arabic-English, Google Translate, machine translation, translation quality assessment

1. Introduction

In the past years, the translation process has substantially changed because of technological advancements, such as the use of Internet and the availability of web-based machine translation (MT) systems (Johnson et al., 2017). MT is an approach to translating texts from one language to another. For a long time, MT had a poor reputation because its output was perceived to be of low quality (e.g., Agarwal et al., 2011). However, recent research has found that the quality of output has improved enough to be used in the translation industry (e.g., Chen, Acosta, & Barry, 2016).

MT has been developed since the 1950s, and different theories and practices have emerged over time. Recently, the quality of neural machine translation (NMT) has been the primary concern of researchers. NMT has emerged as an innovative translation approach that uses deep learning for translation of text in foreign languages (Wu et al., 2016).

Google Translate is one of the most well-known MT systems (Nizke, 2019). It is a free online platform that enables instant translation of documents, words, and sentences. According to Aiken and Ghosh (2009) and Och (2009), Google Translate is frequently used because it provides translation services for several combinations of language pairs and is more accurate than other MT systems, which makes it suitable for this research. Moreover, Google updated its Google Translate system in 2016, moving from statistical machine translation (PBMT) to neural machine translation (NMT). However, based on the results of previous research (e.g., Costa et al., 2015; Kafipour & Jahanshahi, 2015) regarding the challenges encountered by MT, the present study aims to investigate the quality of Google MT for the Arabic-English language pair after it was updated in 2017. Evaluating the quality of Google neural machine translation (GNMT) is a relatively new research field that has not yet been explored extensively (Vardaro et al., 2019).

Evaluating MT is a difficult yet important task mainly because one cannot easily measure the quality of the output. Many correct translations of a text may be possible, whether it was translated by machine or a human translator. One sentence can be translated differently either by several translators or by the same translator.

Depraetere (2011) describes four techniques for evaluating MT: (1) human evaluation of adequacy and fluency, (2) automated evaluation techniques, (3) evaluation based on the analysis of errors, and (4) evaluation based on post-editing time. The present study uses the first and third techniques.

In the MT field, MT can be assessed manually or automatically (Lommel et al., 2014). Although automatic evaluation is objective and cheap, it is less comprehensive than human evaluation (ibid). According to Maučec and Donaj (2019), human evaluation is the most common option for assessing the quality of MT. Hence, automatic evaluation was discarded for this reason.

In human evaluation, the MT output is assessed by expert evaluators, proficient in translation, who should be bilingual in both the source and target languages (ibid). According to Bonnie et al. (2010, p. 809): "The fact is we have no real substitute for human judgments of translations. Such judgments constitute the reference notion of translation quality." Human evaluation can play a crucial role in improving MT technology; hence, research in MT is now moving toward integrating human quality assessment (QA) into the MT field (Girardi, 2014).

In terms of error analysis, this field of study is a part of applied linguistics. It aims to detect problems in translation and reveal the degrees and patterns of errors (Kafipour & Jahanshahi, 2015). In translation, identifying errors is crucial, especially for improving the quality of the end-product (van der Wees, Bisazza, & Monz, 2015). Hence, the present study adopts this method of analysis, wherein we apply (1) human evaluation of adequacy and fluency and (2) human error analysis. In the latter, human evaluators identify and classify translation errors and precisely describe specific deficiencies in the MT output.

In this paper, we provide a detailed overview of the types of errors in GNMT and identify its potential shortcomings using (1) human evaluation of adequacy and fluency and (2) human error analysis methods. Unlike previous works (e.g., Burchardt et al., 2017; Isabelle, Cherry, & Foster, 2017; Oudah et al., 2019), where NMT was compared with PBMT, we take a different approach by examining the possible deficits in the Google Translate system (as the most widely used free engine for translation).

To the best of our knowledge, no previous study has examined the Google Translate output for the English-Arabic language pair using the same methodology—specifically, adopting both (1) human evaluation of adequacy and fluency and (2) error analysis using taxonomies—after Google updated its system in 2017.

2. Google Translate: Statistical and Neural Machine Translation (NMT)

Google Translate was developed in 2006 and launched as the best statistical MT (Och, 2009). The translation process with the use of Google Translate entails using a computer system and the process is based on text patterns rather than using specific language rules as a reference (Turovsky, 2019).

In 2016, Google Translate received a significant update: it was improved by adopting NMT over statistical MT (United Language Group, 2017). NMT is an innovative method of MT, which creates more accurate translations than statistical MTs (Turovsky, 2018). Specifically, NMT uses a neural network—as in the human brain—wherein information is sent to various layers and is processed before the output (Cheng, 2019). NMT mainly focuses on the use of deep learning methods for translating text based on the already developed statistical models (ibid). Moreover, using deep learning techniques allows for faster translations than using statistical models alone. This enhances the ability of NMT to provide a higher-quality output during the translation (Cheng, 2019). Moreover, NMT uses algorithms to provide a better understanding of linguistic rules from the statistical models. One benefit of using NMT is its quality and speed (Cheng, 2019). Thus, NMT is believed to be an essential translation method of the future, and the translation capabilities with the use of NMT will continue to advance. NMT focuses on the translation of a whole sentence at a time (Turovsky, 2018). The current Google Translate is more accurate and has been estimated to be 60 times more accurate than the previous translation system (ibid).

For example, Popović (2018) examined the overall performance of NMT and PBMT for the German–English language pair. She manually annotated 264 sentences for English-to-German and 204 for German-to-English sentences obtained from a corpus of 3000 sentences. She found that the number of correct sentences in NMT was remarkably higher than PBMT. She concluded that NMT outperformed PBMT in terms of verb aspects (form, order, and omission), articles, English noun collocations, and German compounds, as well as phrase structure, which improves fluency. Many other studies have compared the output of NMT and PBMT for many language pairs, including the Arabic language (e.g., Burchardt et al., 2017; Isabelle, Cherry, & Foster, 2017; Oudah et al., 2019). Therefore, we will not be comparing PBMT and NMT in this paper.

NMT was first applied in Google Translate in 2016 to translation between eight languages: English, French, German, Spanish, Portuguese, Chinese, Japanese, Korean, and Turkish (Turovsky, 2017). Later, six more

languages were added in March 2017: Russian, Hindi, Vietnamese, Thai, Hebrew, and Arabic (Jordan, 2017).

3. Related Works

MT can be evaluated by presenting the output of MT to bilingual human evaluators, who understand both source and target languages, to score the quality of a translation (Popovic, 2018). Human evaluators can adopt two different approaches. First, experts can evaluate adequacy (i.e., preservation of meaning) and fluency (i.e., grammaticality and overall quality; based on a combination of both), as well as estimated cognitive post-editing effort. Second, the experts can compare different MTs of the same source text to identify which translation is better without providing any scores (Callison-Burch et al., 2007).

Ghasemi and Hashemian (2016) examined the quality of Google Translate's output for English-Persian and Persian-English translations using MT QA. The study focused on translating 100 selected sentences from Motarjem Harma, an interpreter application. The effectiveness of Google Translate was analyzed based on errors generated by two MT QA systems (Ghasemi & Hashemian, 2016). MT QA was used to analyze the translations using tables for different concepts: wrong word order, errors in the distribution and use of verbs, lexicosemantic errors, and wrong use of tenses. From the results obtained, Ghasemi and Hashemian (2016) found no significant differences between the two systems when translating from English to Persian and from Persian to English. Moreover, the analysis could not identify the error frequency in all types of texts translated by Google Translate.

Several studies have focused on error analysis and classification in the area of MT. Many researchers, such as Llitjós et al. (2005); Vilar et al. (2006); Bojar (2011), focused on design of taxonomies. For example, one of the most referred taxonomies in MT is the classification proposed by Vilar et al. (2006). They extended the work of Llitjós et al. (2005) and classified errors into five categories: "Missing Words," when some words in the translated text (TT) are missing; "Word Order," errors related to word order in the target sentence; "Incorrect Words," errors that occur when the system does not provide the correct translation of a given word; "Unknown Words," words found when the system copies the input word to the TT without changing it; and finally "Punctuation Errors."

Similarly, Vilar et al. (2006) and Bojar (2011) classified errors into four types: "Bad Punctuation," "Missing Word," "Word Order," and "Incorrect Words." Many other studies, such as Popović and Ney (2006), evaluated error identification. In this paper, we examine a linguistically motivated taxonomy for translation errors that extends the previous ones. Our research is different in two ways: first, we provide a detailed examination and analysis of errors in MT output (specifically, for Google Translate) and, second, we examine the quality of MT output in terms of adequacy and fluency.

A related study conducted by Zaghouani (2016) presented guidelines and annotation procedures to create a human-corrected MT corpus for the Modern Standard Arabic. Zaghouani created comprehensive and simplified annotation guidelines with the help of a team of five annotators and one lead annotator. To ensure a high annotation agreement between the annotators, Zaghouani organized several training sessions for the annotators. It was the first published manual post-editing annotation of MT for the English-Arabic language pair (ibid).

Zaghouani created general annotation correction guidelines and classified errors under seven categories: spelling errors (which mostly occur in letters Yaa and Hamza), word choice errors, morphology errors (the use of incorrect inflection or derivation), syntactic errors (gender and number agreement, definiteness, wrong case, and tense assignment), proper name errors (when the names of entities are improperly translated into Arabic), dialectal usage errors (when the dialect is generally not present in the MT texts), and punctuation errors (in some cases, punctuation signs appear in the wrong place).

Bojar (2011) manually identified errors to evaluate four systems: Google Translate, PC Translator12, TectoMT13, and CU-Bojar (Bojar et al., 2009). He applied two techniques of manual evaluation to identify error types discussed in the previously mentioned MT systems. The first technique is "blind post-editing," where the evaluation was performed by two evaluators separately. The first evaluator edited the system output and, thus, produced an edited version. The second evaluator worked on the edited version, compared the source and the reference translation, and judged whether the translation was still acceptable. The second technique was the manual annotation of the errors using a taxonomy inspired by Vilar et al. (2006).

Condon et al. (2010) examined MT English-Iraqi Arabic and vice versa. They classified errors under "Deletions," "Insertions," and "Substitutions" for morphological classes and types of errors, following a similar taxonomy as proposed by Vilar et al. (2006).

No general rules for defining error categories exist (Popovic, 2018). In this paper, we classify errors using a similar approach as previous researchers. However, we use a slightly different taxonomy as the Arabic language

has some specificity. In this regard, Costa et al. (2015, p. 3) affirm that "it is important to say that all taxonomies are influenced by the idiosyncrasies of the languages with which they are working."

4. Translation Quality Assessment (TQA)

Translation QA (TQA) is the process of assessing a translated text in terms of its quality (Munday, 2001). To ensure a valid and reliable assessment, it has to follow particular rules and standards (Williams, 2009). However, the process of determining particular criteria for evaluating translation quality is a difficult task, which is believed to be "probably one of the most controversial intensely debated topics in translation scholarship and practice" (Colina, 2009, p. 236). That is because the assessment criteria are negotiable in the field of translation studies, as the relative nature of quality itself is believed to be too complex and too context dependent to be formulated under one definition (Nord, 1991). However, many researchers agree that assessing translation quality should measure particular issues, such as adequacy and fluency; these two metrics are most commonly used in human evaluation (White, 1994; Callison-Burch, 2007). For example, Gupta et al. (2011) assert that human evaluation is based on adequacy and fluency.

Adequacy (also called accuracy or fidelity) is defined as the extent to which the translation conveys the meaning of the source language unit (Koehn, 2009). Fluency is defined as the extent to which the translation follows the rules and the norms of the target language; thus, it focuses only on the target language unit (Casilho et al. 2018). Importantly, this aspect of evaluating the MT output is normally conducted at the sentence or segment level without considering the context of the translation (ibid).

An error analysis method aims at analyzing errors to obtain an error profile for a translation output (Popovic, 2018). It can be conducted either manually, automatically, or semi-automatically (combined method) (Popovic, 2018). The most obvious method for error analysis is to examine the translation output, mark each error in the translation, and assign a corresponding error tag to it (Guzmán et al., 2015). Error classification aims to identify and classify actual errors in a translated text.

5. Materials and Methods

Although human evaluation is expensive and time consuming, it is more accurate and can provide a more thorough analysis of the errors (Joshi et al., 2015) and can be performed by one or multiple evaluators. In the case of multiple human evaluators, the agreement among them can be calculated to provide additional information on the reliability of the results (ibid).

Data for this study were collected from English articles. We manually examined the samples for readability, potential translation problems, and MT quality. To identify the problems in the output of MT, a deep linguistic error analysis was conducted for a sample of English passages translated into Arabic by GNMT.

A total of 100 English passages were obtained from English articles and were translated into Arabic using Google Translate. The source and target passages were directly compared one by one by human evaluators, who used numerical ranges for judging the quality of the MT output. Specifically, the evaluators used the error analysis method and additionally evaluated adequacy and fluency when examining the GNMT output.

The general process of manual error classification is illustrated in Figure 1. Figure 1 presents error taxonomies that cover both translation aspects and linguistic aspects. Three evaluators, who are experts in linguistics and translation studies for the Arabic-English language pair, conducted a detailed analysis at the translational and linguistic levels and examined the MT translations for adequacy and fluency.

Error analysis at the translation level includes the following seven types of errors. (1) Mistranslation errors (abbreviated in the figure as Mis.) comprise all errors related to incorrect translations of the source language content. (2) "Untranslated" errors (untrans.) occur when the source language content is not translated. (3) "Addition" errors (add.) occur when elements are added to the target text that is not present in the source text. (4) Omission errors (omit.) occur when elements are deleted from the target text that is present in the source text. (5) Lexical errors (lexis.) include word choice errors. (6) Orthographic errors (ortho.) include spelling and punctuation errors, where in some cases punctuation signs appear in wrong places. (7) Miscellaneous errors include errors that do not fall under any of the other categories, such as names of entities or concept that are being improperly translated into Arabic.

At the linguistic level, errors were categorized into three levels: syntactic errors, grammatical errors, and semantic errors. Syntactic errors were subcategorized into errors that occur when the translation starts with a nominal sentence in the place of a verbal sentence in the ST (Nomi. sen. instead of v. sent.) and when the TT violates the entire phrase structure (viol. structure) (e.g., putting adjective before noun). Grammatical errors include violating subject–verb agreement (viol. S–V agree), such as masculine and feminine; singular, dual, and

plural; first, second, and third person; using a noun in place of a verb (N–V); using a verb in place of a noun (V–N); using wrong prepositions, articles, and particles (wrong prep.); using the definite article before genitives (def. before genitives), and omitting functional morphemes (omit. funct.) (i.e., prepositions, articles, conjunctions, pronouns, auxiliary).

Moreover, semantic errors occur when using words with ambiguous meaning (ambiguity); using terms that convey a different meaning (diff. meaning); using unfamiliar words in place of collocations (unfamiliar collo.); using wrong reference and relative pronouns (wrong ref.); adding an unnecessary word, preposition, or article before a word (add. unnecessary words); omitting necessary words or phrases (omit. necessary words); and corrupting the meaning of the entire sentence (corrupt. meaning).

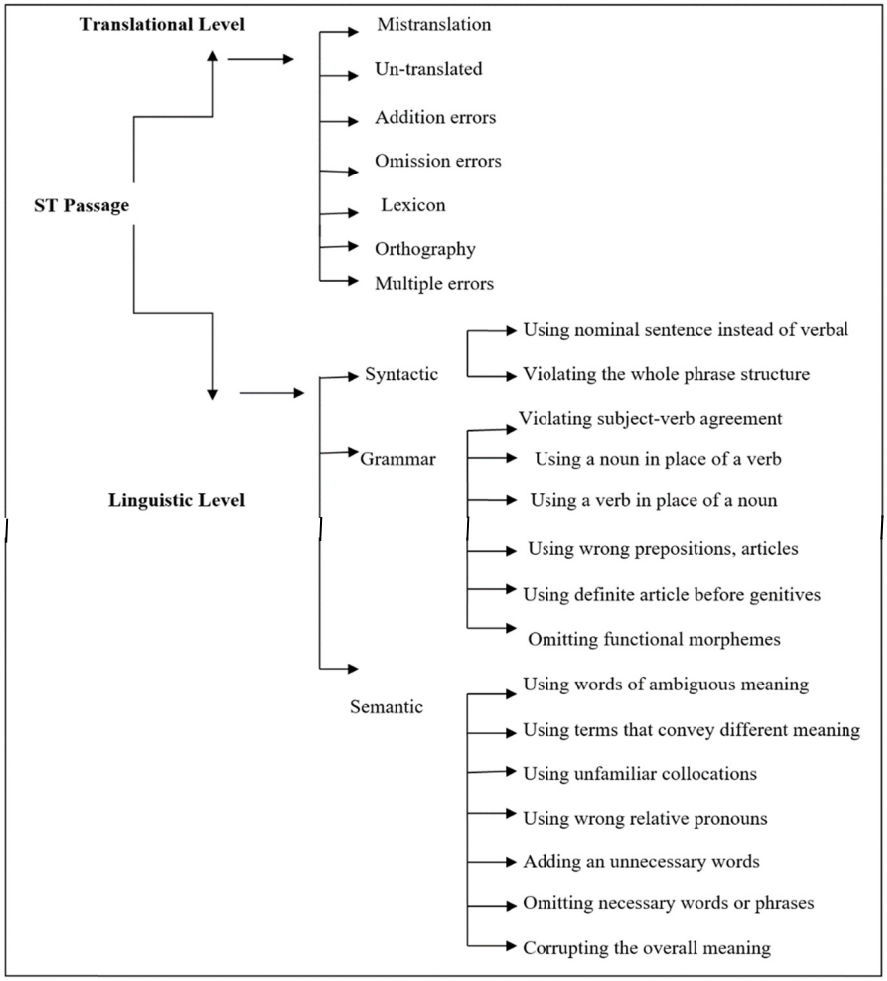

Figure 1. Error taxonomy for the Arabic-English translation used in this study

Evaluators assessed the MT output on a 5-point scale and judged on two aspects: adequacy and fluency. To avoid subjectivity in human evaluation, the three expert evaluators were asked to evaluate the MT output using the evaluation set, and their final evaluations were justified statistically.

Following many studies (White et al., 1994; Gupta et al., 2011; Banchs et al., 2015), human evaluators relied on the following numeric scale for judging adequacy and fluency.

Adequacy can be evaluated by examining both source and target segments to show mainly how much of the source information is preserved. According to White et al. (1994), the adequacy score is defined as follows: "1" is used when none of the meaning is preserved, "2" is used when little of the meaning is preserved, "3" is used when much of the meaning is preserved, "4" is used when most of the meaning is preserved, and "5" is used when all of the meaning is preserved (Table 1).

Conversely, fluency can be evaluated by examining the target segments only; the goal is to examine the language quality of the translated text. The fluency score is defined as follows: "1" is given for incomprehensible target language, "2" is given for a disfluent target language, "3" is given for non-native kind of target language, "4" is given for a good-quality target language, and "5" is given for flawless target language (White et al., 1994).

Table 1. Numeric scale for judging adequacy and fluency

Adequacy		Fluency	
5	All meaning	5	Flawless language
4	Most meaning	4	Good language
3	Much meaning	3	Non-native language
2	Little meaning	2	Disfluent language
1	None	1	Incomprehensible

Evaluators rated the MT output using the predetermined scale described above. The scale ranges from 1 to 5, where 1 is the lowest score and 5 is the highest score.

There are three common inter-rater agreement metrics for the evaluation: the percentage of agreement, various versions of Cohen's kappa measure, and the intra-class correlation coefficient (Graham et al., 2012). The percentage of agreement is the simplest and the most straightforward measure. It provides basic approximation of the evaluators' agreement. Cohen's kappa measure is more rigorous than the percentage of absolute agreement because it considers the evaluators' agreement by chance. Typically, kappa measures the agreement between two raters. The intra-class correlation measures the agreement among evaluators when there are many rating categories (5 or more) or when ratings are made along a continuous scale (ibid).

Evaluators meet multiple times to identify taxonomies and classify the data under different categories and taxonomies. Before evaluating the dataset, evaluators agreed on 19 taxonomies to classify errors in the MT output. Because many errors were identified by one evaluator but not by the others, evaluators had to agree on particular errors to be considered in this analysis. As we have multiple well-defined labels and standards that each evaluator agreed on and clearly understands, the percentage of absolute agreement is used, which simply calculates the number of times evaluators agree on a rating. Importantly, evaluators have undergone training to develop a common understanding of how to apply the rating system as consistently as possible. Previous research shows that such a training improves accuracy, reliability, and validity (Woehr & Huffcutt, 1994; Gorman & Rentsch, 2009; etc.).

Subsequently, the inter-evaluator agreement was calculated for each label separately based on the evaluators' decisions at the meeting. The same approach was used for the adequacy and fluency measures. Their agreement was calculated for each label, and the average scores are presented in Table 2.

To judge whether an inter-rater agreement is sufficient or not, various experts (e.g., Hartmann, 1977; Stemler, 2004) contend that when using the percentage of absolute agreement, values from 75% to 90% demonstrate an acceptable level of agreement.

6. Results and Discussion

All of the evaluators identified and classified the errors at the sentence level in 100 passages translated by MT. Evaluators' agreement was first compared in terms of error localization to ensure that all evaluators agree whether there is an error in the sentence or not. Then, we took all agreed errors for all 19 classifications and added them to a separate column for better visualization of the results. In other words, errors must be agreed upon all evaluators to be considered as an error. Once the data were evaluated, the inter-rater agreement was calculated using the percentage of absolute agreement.

6.1 Error Taxonomies

As shown in Figure 2, evaluator 1 identified that mistranslation errors were the most common in the MT output. The second most common type was "corrupting the overall meaning of the sentence" followed by "lexical errors." Omitting necessary words category had zero errors and thus it is the lowest percentage in all categories. Figure 2 also shows that the least frequent errors were related to using definite articles before genitives, using unfamiliar words in place of collocations, using terms that convey very different meaning and using a noun in place of a verb.

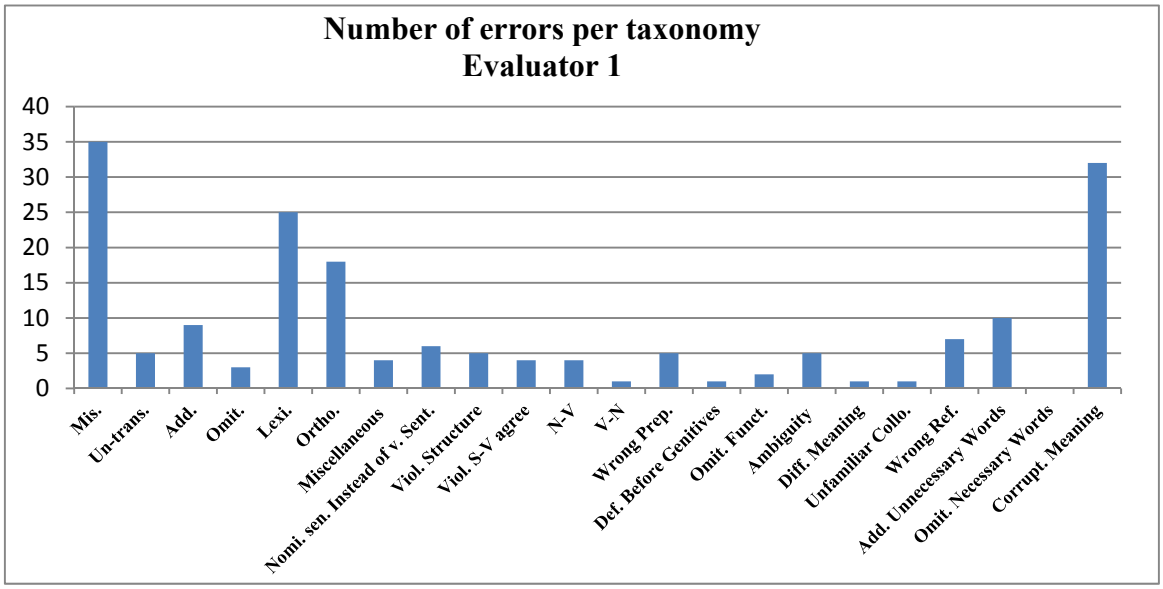

Figure 2. Number of errors per taxonomy for evaluator 1

Figure 3 shows the results for evaluator 2. The numbers suggest that mistranslation errors were the most common, followed by corrupting the overall meaning of the sentence and orthography. Moreover, using the definite article before genitives, wrong references, and omitting necessary words had the lowest percentage in all categories.

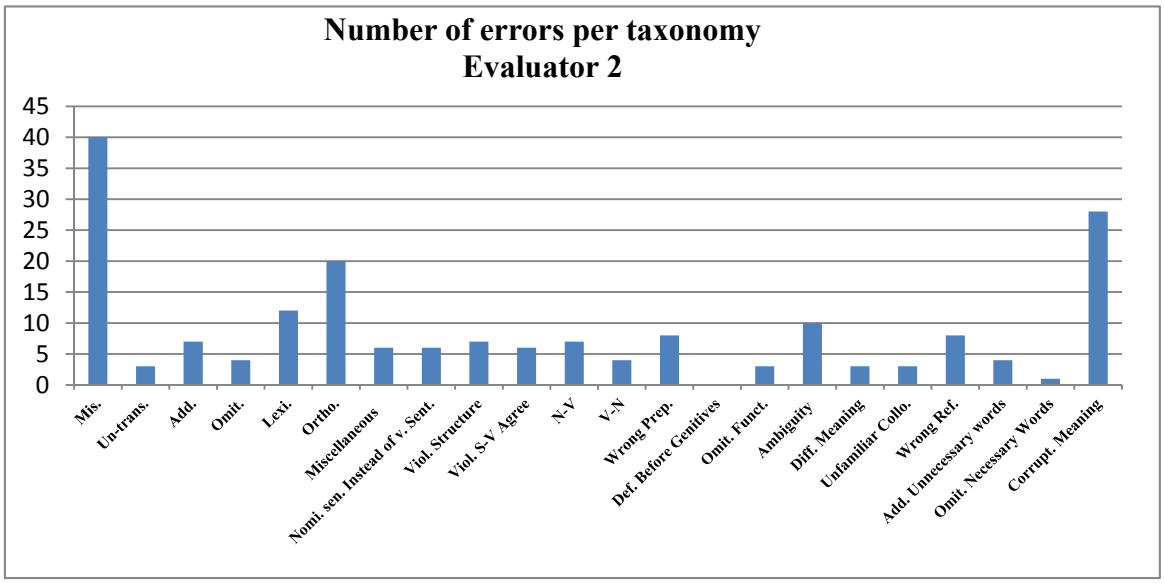

Figure 3. Number of errors per taxonomy for evaluator 2

Figure 4 shows the results for evaluator 3. Here, mistranslation errors were the most common, followed by orthographic errors and corrupting the meaning of the sentence. On the contrary, errors as a result of using definite articles before genitives and omitting necessary words or phrases had the lowest frequency.

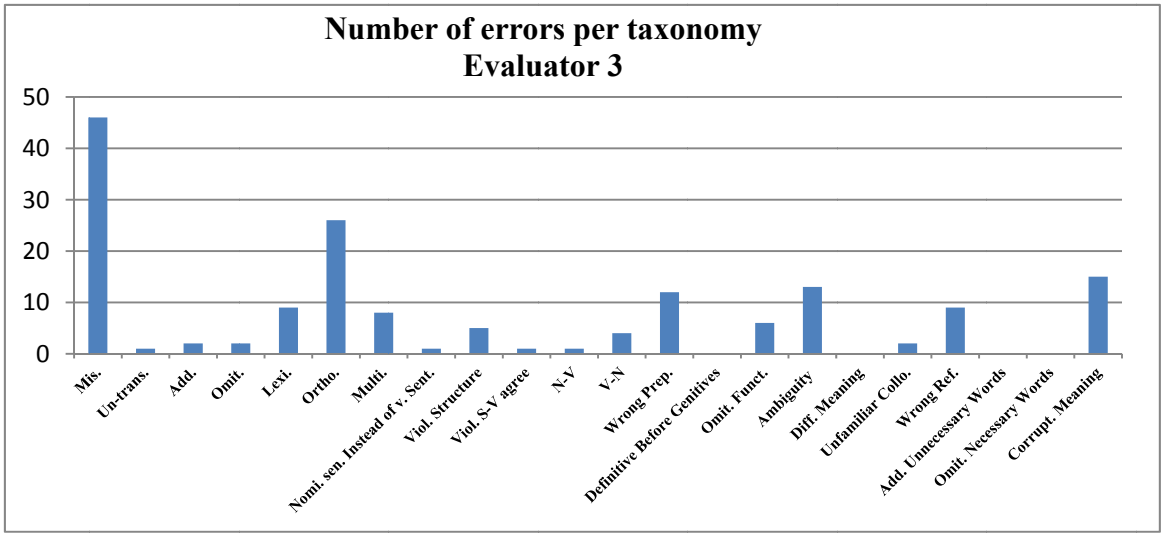

Figure 4. Number of errors per taxonomy for evaluator 3

In general, Figure 5 shows that the three evaluators agreed on most of the errors. The scores are relatively consistent across all error types. Thus, according to many experts (Hartmann, 1977; Stemler, 2004), these results constitute an acceptable level of agreement.

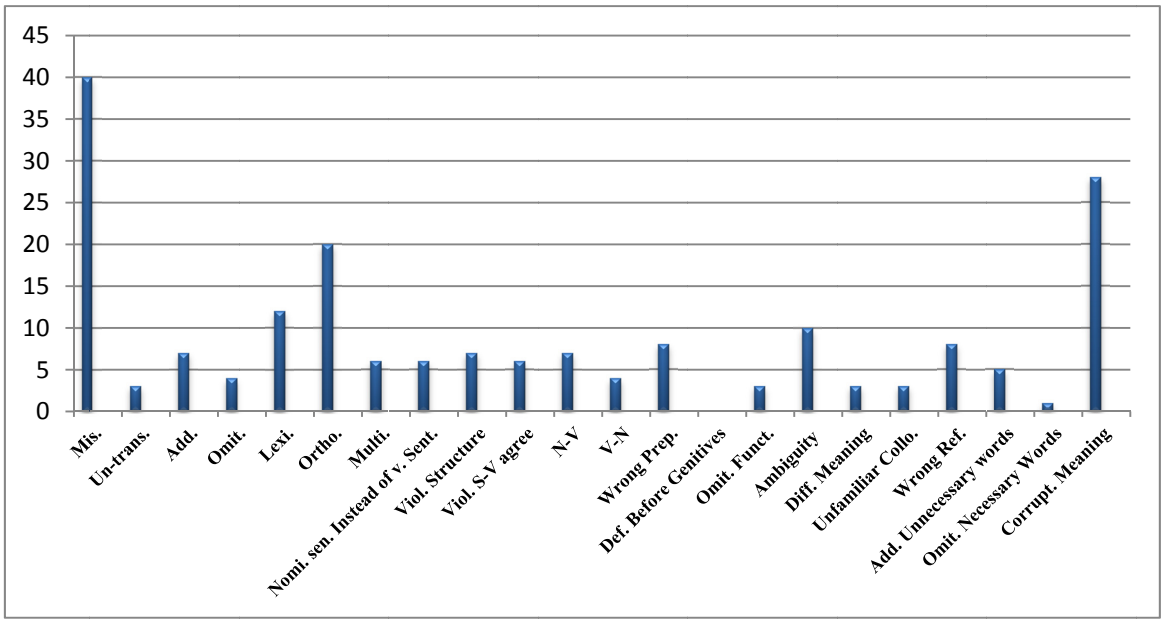

Figure 5. Inter-rater agreement per taxonomies

6.1.1 Translation Level

At the translation level, we consider all errors related to the way each word is translated: mistranslated, deleted, or translated by addition. Table 2 shows the number of errors by each evaluator and the number of agreed errors.

Table 2. Number of errors and the agreed errors by the evaluators at the translation level

Translation level	Number of agreed errors	EVAL 1	EVAL 2	EVAL 3
Mistranslation	40	35	40	46
Addition	4	9	7	2
Omission	2	3	4	2

To illustrate how evaluators identified and classified errors in the data, this section gives one example for each of the 19 classifications.

1) Example of a mistranslation error

English (EN): The Barcelona verdict comes despite a ruling in a similar case by the Spanish Supreme Court upgrading a conviction from sexual abuse to sexual assault.

Arabic (ARB):

ويأتي قرار برشلونة على الرغم من صدور حكم في قضية مماثلة من قبل المحكمة العليا الإسبانية بتطوير الإدانة من الاعتداء الجنسي إلى الاعتداء الجنسي.

Correct translation:

ويأتي قرار برشلونة على الرغم من صدور حكم في قضية مماثلة من قبل المحكمة العليا الإسبانية بتطوير الإدانة من العنف الجنسي إلى الاعتداء الجنسي.

The expression "sexual abuse" has been mistranslated by MT. It could not differentiate between the expressions "sexual abuse" and "sexual assault," as they have been translated identically. However, there is a difference between the words "abuse" and "assault." The correct translation maintains this difference.

2) Example of an omission error

English (EN): The US president also sparred at the White House with a Reuters correspondent, who asked him what he considered treasonous.

Arabic (ARB):

كما انتشر الرئيس الأمريكي في البيت الأبيض مع مراسل لرويترز ، وسألوه عما اعتبره خيانة.

Correct translation:

كما تشاجر الرئيس الأمريكي في البيت الأبيض مع مراسل لرويترز ، وسألوه عما اعتبره خيانة.

In this example, two errors occurred. The word "sparred" has been omitted from the TT. Moreover, this omission entails mistranslation: the ST word has been mistranslated as انتشر (i.e., "spread"). Evaluators agreed to insert this error as both omission and mistranslation.

3) Example of an addition error

English (EN): Starting early Wednesday, crowds gathered in a half-dozen neighborhoods across Baghdad, with riot police attempting to disperse them using tear gas and firing live rounds into the air.

Arabic (ARB):

ابتداءً من صباح الأربعاء ، تجمعت الحشود في نصف حي من أحياء بغداد ، حيث حاولت شرطة مكافحة الشغب تفريقهم باستخدام الغاز المسيل للدموع وإطلاق الرصاص الحي في الهواء.

Correct translation:

ابتداءً من صباح الأربعاء ، تجمعت الحشود في ستة أحياء في بغداد ، حيث حاولت شرطة مكافحة الشغب تفريقهم باستخدام الغاز المسيل للدموع وإطلاق الرصاص الحي في الهواء.

In the above example, MT has added the word حي to the TT, which does not exist in the ST. Moreover, mistranslation errors occur in this example.

6.1.2 Linguistic Level

At the linguistic level, the evaluators considered fluency errors, which affected the quality of writing in the target language. This included lexical errors, orthographic errors, and miscellaneous error when errors do not fall under any of the other categories. Moreover, all grammar, syntactic, and semantic errors were identified in the MT output.

As shown in Table 3, the three evaluators agreed on 17 lexical errors, 26 orthographic errors, and 7 miscellaneous errors. Examples of these errors are listed below.

Table 3. Number of errors and the errors agreed by the evaluators at the linguistic level

Linguistic level	Number of agreed errors	EVAL 1	EVAL 2	EVAL 3
Lexical errors	17	25	12	9
Orthographic errors	26	18	20	26
Miscellaneous error	7	4	6	8

1) Example of a lexical error

English (EN): Ms. Masala has already confessed to killing Marilyne Planche, 52, during a fight at the victim's apartment.

Arabic (ARB):

اعترفت ماسالا بالفعل بقتل مارلين بلانش ، 52 عامًا ، خلال معركة في شقة الضحية.

Correct translation:

اعترفت ماسالا بالفعل بقتل مارلين بلانش ، 52 عامًا ، خلال شجار في شقة الضحية.

MT has translated the word "fight" as معركة. However, the TT word is used inappropriately in this context, as the word معركة refers to a fight in a battle. Such inappropriate usages of words are identified as errors in the MT output.

2) Example of an orthographic error

English (EN): A parade celebrating the formal ascension of Japan's Emperor Naruhito has been postponed in the wake of Typhoon Hagibis. The parade, which sees the emperor travel in an open-top car to "meet" the public, was postponed out of respect for the victims and their families.

Arabic (ARB):

تم تأجيل عرض يحتفل بالصعود الرسمي للإمبراطور الياباني ناروهيتو في أعقاب إعصار هاجيبيس. تم تأجيل العرض ، الذي يرى أن الإمبراطور يسافر في سيارة مكشوفة "للقاء" الجمهور ، احتراماً للضحايا وعائلاتهم.

Correct translation:

تم تأجيل عرض يحتفل بالصعود الرسمي للإمبراطور الياباني ناروهيتو في أعقاب إعصار هاجيبيس. تم تأجيل العرض ، الذي يرى أن الإمبراطور يسافر في سيارة مكشوفة "للقاء" الجمهور ، احتراماً للضحايا وعائلاتهم.

Orthographic errors include punctuation, capitalization, and spelling errors. In this example, the underlined sentence is nonessential information that is added parenthetically to a sentence; it is separated from the main sentence by commas before and after the sentence. The MT replicated the same punctuation system of the English language in the TT. However, the Arabic language does not have a parenthetical phrase or sentence; thus, commas are used wrongly in this situation. Moreover, the Arabic language does not have capitalization; hence, this category is discarded from the analysis.

3) Example of miscellaneous error

English (EN): Over a hundred demonstrators were arrested at yellow vest protests in Paris on Saturday as about 7,500 police were deployed to deal with the movement's radical anarchist "black blocs" strand.

Arabic (ARB):

تم إلقاء القبض على أكثر من مائة متظاهر في مظاهرات بالسترات الصفراء في باريس يوم السبت ، حيث تم نشر حوالي 7500 شرطي للتعامل مع "الكتل السوداء" الأناركية المتطرفة للحركة.

Correct translation:

تم إلقاء القبض على أكثر من مائة متظاهر في مظاهرات بالسترات الصفراء في باريس يوم السبت ، حيث تم نشر حوالي 7500 شرطي للتعامل مع "الكتل السوداء" اللاسلطوية المتطرفة للحركة.

Miscellaneous errors are related to different types of errors such as the word "anarchist". It refers to a person who rebels against authority. This word has been translated as أناركية using the transliteration strategy. However, as the word "anarchist" has a direct equivalent in the Arabic language اللاسلطوية, one could argue that the use of transliteration strategy is not the best option.

Similarly, evaluators agreed on four errors of starting with a nominal sentence in place of a verbal one and agreed on five errors related to violating the entire phrase structure. Table 4 demonstrates the numbers of identified errors separately for each evaluator and the agreed number of errors.

Table 4. Number of errors and the errors agreed by the evaluators at the syntactic level

Syntactic errors	Number of agreed errors	EVAL 1	EVAL 2	EVAL 3
Starting with a nominal sentence in the place of a verbal sentence.	4	6	6	1
Violating the whole phrase structure (e.g., putting adjective before noun).	5	5	7	5

4) Example of an error related to starting with a nominal sentence in the place of a verbal sentence

English (EN): European powers, while criticising Iran, want to salvage a 2015 accord under which Iran dramatically scaled back its nuclear programme in exchange for unmet promises of sanctions relief.

Arabic (ARB):

القوى الأوروبية ، رغم انتقادها لإيران ، تريد إنقاذ اتفاق عام 2015 الذي بموجبه قامت إيران بتخفيض برنامجها النووي بشكل درامي في مقابل وعود غير مستوفاة بتخفيف العقوبات.

Correct translation:

تريد القوى الأوروبية برغم انتقادها لإيران إنقاذ اتفاق عام 2015 الذي بموجبه قامت إيران بتخفيض برنامجها النووي بشكل درامي في مقابل وعود غير مستوفاة بتخفيف العقوبات.

In English grammar, the sentence should always starts with a subject. However, this is not the case in Arabic grammar. In this example, the Arabic translation of the sentence followed the same word order (subject or noun + verb) of the English structure instead of following the Arabic grammar (i.e., using a verbal sentence).

5) Example of an error related to violating the entire phrase structure

English (EN): The wave of arrests comes ahead of a "million-man march" Friday called for by an exiled businessman whose online videos accusing Sisi and the military of corruption sparked last week's rallies.

Arabic (ARB):

وتأتي موجة الاعتقالات قبل "مسيرة مليون رجل" دعا إليها يوم الجمعة رجل أعمال منفي شجبت مقاطع فيديو على الإنترنت تتهم السيسي والجيش بالفساد مسيرات الأسبوع الماضي.

Correct translation:

وتأتي موجة الاعتقالات قبل "مسيرة مليون رجل" دعا إليها يوم الجمعة رجل أعمال منفي شجبت مقاطع فيديو على الإنترنت تتهم السيسي والجيش بالفساد الذي أشعل الشرارة لانطلاق مسيرات الأسبوع الماضي.

The MT sentence is not comprehensible because of a problem in its structure: the verb is omitted from the Arabic translation, which makes the sentence difficult to understand.

In terms of grammar, as demonstrated in Table 5, evaluator 1 identified four errors related to violating the subject–verb agreement, evaluator 2 identified six, and evaluator 3 identified one. However, at the meeting, evaluators agreed on three errors only. The case applies to the rest of the errors, as they agreed on 3 errors related to using a noun in place of a verb, 3 errors with using a verb in place of a noun, 10 errors with using wrong preposition or articles, and 1 error with using the definite article before genitives.

Table 5. Number of errors and the errors agreed by the evaluators at the grammatical level

Grammatical Errors	Number of agreed errors	EVAL 1	EVAL 2	EVAL 3
Violating the subject–verb agreement (masculine and feminine; singular, dual, and plural; first, second, and third person)	3	4	6	1
Using a noun in place of a verb	3	4	7	1
Using a verb in place of a noun	3	1	4	4
Using wrong prepositions, articles, and particles	10	5	8	12
Using definite articles before genitives	1	1	0	0

6) Example of an error related to violating the subject-verb agreement

English (EN): The leaders said resolving the conflict is the only way to ensure peace in the region, urging the international community to take action to put a stop to the building and expansion of illegal settlements.

Arabic (ARB):

وقال الزعماء إن حل النزاع هو الطريقة الوحيدة لضمان السلام في المنطقة ، وحث المجتمع الدولي على اتخاذ إجراءات لوقف بناء وتوسيع المستوطنات غير القانونية.

Correct translation:

وقال الزعماء إن حل النزاع هو الطريقة الوحيدة لضمان السلام في المنطقة ، وحثوا المجتمع الدولي على اتخاذ إجراءات لوقف بناء وتوسيع المستوطنات غير القانونية.

In Arabic grammar, the subject should agree with the verb in terms of gender and number. The word حث is a singular verb that does not agree with its subject "leaders." The plural suffix "*واو*" and "*ا*" should be added to the

verb حث to agree with its subject.

7) Example of an error related to using a noun in place of a verb

English (EN): We need to get back 0 to have frank and demanding discussions on Iran's nuclear, regional and ballistic activities but also to have a broader approach than sanctions.

Arabic (ARB):

نحتاج إلى العودة حول الطاولة لنجري مناقشات صريحة ونطالب بمناقشات بشأن الأنشطة النووية والإقليمية والباليستية لإيران ولكن أيضا لنهج أوسع من العقوبات.

Correct translation:

نحتاج إلى العودة الى الطاولة لنجري مناقشات صريحة ونطالب بمناقشات بشأن الأنشطة النووية والإقليمية والباليستية لإيران ونطالب بنهج أوسع من العقوبات.

The Arabic translation is not clear, as the sentence starts directly with a noun without prior information about it. In this particular situation, a verb should be added to the Arabic sentence to clarify the meaning.

8) Example of an error related to using a verb in place of a noun

English (EN): Protesters—many of them high school and university students—jumped turnstiles, attacked several underground stations, started fires and blocked traffic, leaving widespread damage across the city and thousands of commuters without transport.

Arabic (ARB):

قفز المتظاهرون - وكثير منهم من طلاب المدارس الثانوية والجامعات - الباب الدوار ، وهاجموا العديد من محطات المترو ، وبدأوا في إطلاق النار وحظرت حركة المرور ، وترك أضرار واسعة النطاق في جميع أنحاء المدينة وآلاف الركاب دون وسائل النقل.

Correct translation:

قفز المتظاهرون - وكثير منهم من طلاب المدارس الثانوية والجامعات - الباب الدوار ، وهاجموا العديد من محطات المترو ، وبدأوا في إطلاق النار واعاقة حركة المرور مما خلف أضرارا واسعة النطاق في جميع أنحاء المدينة وآلاف الركاب دون وسائل النقل.

The word "leaving" is translated as a singular verb *ترك* , which distorts the meaning of the Arabic sentence, as it does not have any clear subject.

9) Example of an error related to using wrong prepositions, articles, and particles

English (EN): The trade agreement did not mention car tariffs of up to 25%, which were previously threatened by the US.

Arabic (ARB):

لم تذكر الاتفاقية التجارية تعريفات السيارات بنسبة تصل إلى 25 ٪ ، والتي كانت مهددة من قبل من قبل الولايات المتحدة.

Correct translation:

لم تذكر الاتفاقية التجارية تعريفات السيارات بنسبة تصل إلى 25 ٪ ، والتي كانت مهددة سابقا من قبل الولايات المتحدة.

The adverb "previously" has been translated as *من قبل*, and this not the correct translation of this adverb, especially when both words "previously" and "by" were translated the same as *من قبل* in the same sentence.

10) Example of an error related to using the definite article before genitives

English (EN): The Australian town of Kingaroy in Queensland was hit by a fierce dust storm on Thursday, with winds reaching up to 90km/h (56 mph).

Arabic (ARB):

عرضت مدينة كينجاروي الأسترالية في ولاية كوينزلاند للعاصفة الترابية الشديدة يوم الخميس ، حيث وصلت سرعة الرياح إلى 90 كم / ساعة 56 ميلاً في الساعة.

Correct translation:

عرضت مدينة كينجاروي الأسترالية في ولاية كوينزلاند لعاصفة ترابية شديدة يوم الخميس ، حيث وصلت سرعة الرياح إلى 90 كم / ساعة 56 ميلاً في الساعة.

The indefinite English article "a" is translated as a definite article in Arabic using the prefix "*ال*". This translation corrupts the structure and the meaning of the Arabic translation.

Finally, Table 6 shows the evaluators' agreement at the semantic level. They agreed on 11 errors with using ambiguous words, 1 error with using terms of different meaning, 1 error with incorrect collocations, 9 errors with using a wrong reference, 4 errors with adding unnecessary words, 0 errors with omitting necessary words, and 20

errors with corrupting the meaning of the entire sentence.

Table 6. Number of errors and the agreed errors between the evaluators at the semantic level

Semantic Errors	Number of agreed errors	EVAL 1	EVAL 2	EVAL 3
Using words of ambiguous meaning	11	5	10	13
Using terms that convey very different meaning	1	1	3	0
Using unfamiliar words in place of collocations	1	1	3	2
Using wrong reference and relative pronouns	9	7	8	9
Adding an unnecessary word, preposition, or article before a word	4	10	4	0
Omitting necessary words or phrases	0	0	1	0
Corrupting the meaning of the whole sentence	20	32	28	15

11) Example of an error related to ambiguous words

English (EN): Ministers say the suspension, or prorogation, is not a court matter, but critics argue it was intended to limit scrutiny of the PM's Brexit plans.

Arabic (ARB):

قول الوزراء إن التعليق ، أو الاختصاص ، ليس مسألة محكمة ، لكن النقاد يقولون إن القصد منه هو الحد من التدقيق في خطط رئيس الوزراء بريكسيت.

Correct translation:

قول الوزراء إن التعليق ، أو الاختصاص ، ليس مسألة محكمة ، لكن النقاد يقولون إن القصد منه هو الحد من التدقيق في خطط رئيس الوزراء في خروج المملكة المتحدة من الاتحاد الأوروبي.

The term "Brexit" refers to the withdrawal of the UK from the European Union. In the Arabic translation, the word has been transliterated without any explanation.

12) Example of an error related to using the terms that may convey a different meaning

English (EN): Johnson wants to keep this date, but many MPs fear his threat to leave without agreeing divorce terms with Brussels would cause huge disruption.

Arabic (ARB):

يريد جونسون الحفاظ على هذا الموعد ، لكن العديد من النواب يخشون من أن تهديده بالمغادرة دون الموافقة على شروط الطلاق مع بروكسل قد يتسبب في اضطراب كبير.

Correct translation:

يريد جونسون الحفاظ على هذا الموعد ، لكن العديد من النواب يخشون من أن تهديده بالمغادرة دون الموافقة على شروط الانفصال عن بروكسل قد يتسبب في اضطراب كبير.

The word "divorce" has been translated to Arabic literally as "ending up a marriage." However, the word *إنفصال*, which means "separation," is more appropriate in this context.

13) Example of an error related to using unfamiliar words in place of collocations

English (EN): US President Donald Trump has lashed out at congressional Democrats after they vowed to summons the White House to produce documents this week.

Arabic (ARB):

انتقد الرئيس الأمريكي دونالد ترامب الديمقراطيين في الكونجرس بعد أن تعهدوا باستدعاء البيت الأبيض وثائق هذا الأسبوع.

Correct translation:

انتقد الرئيس الأمريكي دونالد ترامب الديمقراطيين في الكونجرس بعد أن تعهدوا بتقديم البيت الأبيض لوثائق هذا الأسبوع.

Although the words "produce" and "documents" collocate with each other in the English language, they do not collocate in Arabic. Therefore, a better collocation should be used in Arabic to achieve idiomaticity.

14) Example of an error related to using wrong reference and relative pronouns

English (EN): The missile—which was able to carry a nuclear weapon—was the North's 11th test this year. But this one, fired from a platform at sea, was capable of being launched from a submarine.

Arabic (ARB):

كان الصاروخ - الذي كان قادرا على حمل سلاح نووي - هو اختبار كوريا الشمالية الحادي عشر هذا العام. ولكن هذا واحد ، أطلق من منصة في

البحر ، وكان من الممكن إطلاقها من غواصة.

Correct translation:

كان الصاروخ - الذي كان قادرا على حمل سلاح نووي - هو اختبار كوريا الشمالية الحادي عشر هذا العام. ولكن هذا الصاروخ، أطلق من منصة في البحر ، وكان من الممكن إطلاقها من غواصة.

The indicative article "this" refers to "missile" in the English sentence. However, "this" refers to number one instead of "missile" in the Arabic translation.

15) Example of an error related to adding an unnecessary word, preposition, or article before a word

English (EN): The tech firm had been supported by Microsoft, Wikipedia's owner the Wikimedia Foundation, the non-profit Reporters Committee for Freedom of the Press, and the UK freedom of expression campaign group Article 19, among others.

Arabic (ARB):

حصلت شركة التكنولوجيا على دعم من Microsoft *ومؤسسة* Wikipedia *ومالك* Wikipedia *ولجنة مراسلون من أجل حرية الصحافة غير الربحية ومجموعة حملة حرية التعبير في المملكة المتحدة ، المادة 19 ، من بين آخرين.*

Correct translation:

حصلت شركة التكنولوجيا على دعم من شركة مايكروسوفت ومالك مؤسسة ويكيبيديا ولجنة مراسلون من أجل حرية الصحافة غير الربحية ومجموعة حملة حرية التعبير في المملكة المتحدة ، المادة 19 ، من بين آخرين.

The repetition of the word "Wikipedia" is unnecessary in the sentence and confuses the reader.

16) Example of corrupting the meaning of the whole sentence

English (EN): The narrative effectively folds Trump's apparent transgression into an extension of the effective 2016 campaign pitch that only a rule breaker can crush the power of the Washington swamp.

Arabic (ARB):

تطوي الرواية فعليًا تجاوزات ترامب الواضحة إلى امتداد لمرحلة الحملة الفعالة لعام 2016 والتي لا يمكن إلا لسحق القواعد أن يسحق قوة مستنقع واشنطن.

Correct translation:

تطوي الرواية فعليًا تجاوزات ترامب الواضحة إلى امتداد لمرحلة الحملة الفعالة لعام 2016 والتي لا يمكن إلا لمن يخرج عن القاعدة أن يسحق قوة الفساد في واشنطن.

The phrase "Washington swamp" is a metaphor used by politicians in the US to refer to corruption. This phrase has been translated literally, producing a meaningless phrase in Arabic.

In conclusion, the above tables show the number of errors by each evaluator and present the errors that evaluators agreed on. For example, in terms of mistranslation, evaluator 1 identified 35 errors in the MT output, evaluator 2 identified 40 errors, and evaluator 3 identified 46 errors. After the three evaluators discussed and shared their evaluation of the data, they agreed on 40 errors, as shown in Table 2 (first column).

Different types of errors received a different amount of agreement. For instance, the evaluators declared that orthographical errors can be detected easily, as it was easy for them to identify the location of the error; as a result, they easily agreed on 26 errors. However, the evaluators did not agree greatly on the category "adding an unnecessary word, preposition, or article before a word," as they found it hard to decide which words are unnecessary. In this case, evaluator 1 identified 10 errors, evaluator 2 identified 4 errors, and evaluator 3 identified 0 errors. However, they only agreed on four errors in the MT output.

6.2 QA of Adequacy and Fluency

Using the same dataset, we calculated the evaluators' QA of adequacy and fluency on a 5-point scale. After analyzing the data, the overall statistics in Table 2 shows the average adequacy and fluency scores for each evaluator. Hence, we show the average scores given by each of the three evaluators for adequacy and fluency in their evaluation of the Google Translator output from English to Arabic.

Table 7. Evaluator 1 QA of adequacy and fluency

Evaluator 1	Adequacy	Fluency
Total	349	348
Average (1–5 scale)	3.49	3.48
Percentage (100%)	0.70	0.70

Table 8. Evaluator 2 QA of adequacy and fluency

Evaluator 2	Adequacy	Fluency
Total	343	419
Average (1–5 scale)	3.43	4.19
Percentage (100%)	0.69	0.84

Table 9. Evaluator 3 QA of adequacy and fluency

Evaluator 3	Adequacy	Fluency
Total	353	379
Average (1–5 scale)	3.53	3.79
Percentage (100%)	0.71	0.76

Evaluation of adequacy in the translation from English to Arabic showed an excellent consistency, as the three evaluators provided scores in the range of 69–70–71 with an average score of approximately 70%. Similarly, the three evaluators exhibited a reasonable amount of consistency in terms of the evaluation fluency, as they provided scores in the range of 70–76–84 with an average score of approximately 77%, as shown in Table 10.

Table 10. Evaluators QA scores for English-Arabic translations

Evaluators	Adequacy	Fluency
Evaluator 1	70/100	70/100
Evaluator 2	69/100	84/100
Evaluator 3	71/100	76/100
Final score	**70/100**	**77/100**

According to our results, we conclude that the most dominant errors in the MT output were mistranslation errors, followed by corruption of the overall meaning of the sentence and then orthographic errors. In terms of the QA of adequacy and fluency, the results were 70% for accuracy and 77% for fluency. Therefore, according to these results for English-to-Arabic translation, Google Translate produces sentences with relatively few errors, and the translated text is fluent to some extent.

7. Conclusion

In this study, we have conducted a fine-grained manual evaluation to identify and present the dominant types of translation errors produced by Google Translate. The final results suggest that the existing errors in the MT output are mainly related to mistranslations, corruption of the overall meaning of a sentence, and orthographic errors. Moreover, according to the results of our evaluation, Google Translate produces sentences with relatively few errors in English-to-Arabic translation, and the translated text is fluent to some extent. These results can help other researchers in the field to examine these three types of errors more closely and, thus, explain the reason behind the failure in translation at these three levels. From an information technology perspective, it seems that there is a need to develop a more intelligent translation software that considers the context of texts in the translation process. Also, further research is needed to complement the findings of the current one; the use of MT in translating specialized texts might show different weaknesses. Finally, we believe our empirical findings represent a significant contribution to the field of evaluating and improving Google Translate if the current results of errors analysis for Arabic- English languages are taken into consideration.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. *ACL HLT*, 30.

Aiken, M., & Ghosh, K. (2009). Automatic translation in multilingual business meetings. *Industrial Management & Data Systems*, *109*(7), 916–925. https://doi.org/10.1108/02635570910982274

Banchs, R. E., D'Haro, L. F., & Li, H. (2015). Adequacy—fluency metrics: Evaluating MT in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *23*(3), 472–482. https://doi.org/10.1109/TASLP.2015.2405751

Bojar, O. (2011). Analyzing error types in English-Czech machine translation. *Prague Bulletin of Mathematical Linguistics*, *95*, 63–76. https://doi.org/10.2478/v10108-011-0005-2

Bojar, O., Mareˇcek, D., Nov´ak, V., Popel, M., Pt´aˇcek, J., Rouˇs, J., & ˇZabokrtsk´y, Z. (2009). English-Czech MT in 2008. In Proceedings of the Fourth Workshop on Statistical Machine Translation. *Association for Computational Linguistics*, 125–129. https://doi.org/10.3115/1626431.1626457

Bonnie, D., Snover, M., & Madnani, N. (2010). *Machine translation evaluation* (Chapter 5). Retrieved August 22, 2019, from https://www.semanticscholar.org/paper/Part-5%3A-Machine-Translation-Evaluation-Chapter-5.1-Dorr-Snover/22b6bd1980ea40d0f095a151101ea880fae33049

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.-T., & Williams, P. (2017). A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 159–170. https://doi.org/10.1515/pralin-2017-0017

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). *(Meta-)evaluation of machine translation* (pp. 136–158). Association for Computational Linguistics, Proceedings of the Second Workshop on Statistical Machine Translation. Prague, Czech Republic, June. https://doi.org/10.3115/1626355.1626373

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In J. Moorkens, S. Castilho, F. Gaspari & S. Doherty (Eds.), *Translation Quality Assessment from Principles to Practice*. Springer: Switzerland. https://doi.org/10.1007/978-3-319-91241-7_2

Chen, X., Sandra, A., & Adam, B. (2016). Evaluating the accuracy of Google Translate for diabetes education material. *JMIR Diabetes*, *1*, e3. https://doi.org/10.2196/diabetes.5848

Cheng, Y. (2019). *Joint Training for Neural Machine Translation*. Singapore: Springer Nature Singapore Ltd. https://doi.org/10.1007/978-981-32-9748-7

Colina, S. (2009). Further evidence for a functionalist approach to translation quality evaluation. *Target*, *21*, 235–264. https://doi.org/10.1075/target.21.2.02col

Condon, S. L., Parvaz, D., Aberdeen, J. S., Doran, C., Freeman, A., & Awad, M. (2010). *Evaluation of machine translation errors in English and Iraqi Arabic* (pp. 159–168). LREC, European Language Resources Association. https://doi.org/10.21236/ADA576234

Costa, A., Wang, L., Tiago, L., Rui, C., &Luisa, C. (2015). A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation*, *29*, 127–161. https://doi.org/10.1007/s10590-015-9169-0

Girardi, C., Bentivogli, L., Farajian, M., & Federico, M. (2014). *MTEQuAl: A Toolkit for human assessment of machine translation output* (pp. 120–123). COLING Dublin City University and ACL. Retrieved from http://www.aclweb.org/anthology/C14-2026

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology*, *94*(5), 1336–1344. https://doi.org/10.1037/a0016476

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Center for Educator Compensation and Reform (CECR): USA.

Gupta, V., Joshi, N., & Mathur, I. (2011). *Evaluation of English to Urdu machine translation*. Springer-Verlag Berlin Heidelberg. Retrieved September 22, 2019, from https://www.researchgate.net/profile/Vaishali_Gupta7/publication/268508392_Evaluation_of_English_to_Urdu_Machine_Translation/links/546d8f070cf26e95bc3cb6b6/Evaluation-of-English-to-Urdu-Machine-Translation.pdf

Guzman, F., Abdelali, A., Temnikova, I., Sajjad, H., & Vogel, S. (2015). *How do Humans Evaluate Machine Translation* (pp. 457–466). Proceedings of the Tenth Workshop on Statistical Machine Translation. https://doi.org/10.18653/v1/W15-3059

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability measures. *Journal of Applied Behavior Analysis*, *10*, 103–116. https://doi.org/10.1901/jaba.1977.10-103

Isabelle, P., Cherry, C., & Foster, G. (2017). *A challenge set approach to evaluating machine translation* (pp. 2486–2496). Association for Computational Linguistics, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, Denmark, September. https://doi.org/10.18653/v1/D17-1263

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., … Dean, J. (2017). Google's multilingual neural machine translation system: enabling zero-shot translation. *Computation and Language*, *1*. https://arxiv.org/abs/1611.04558v2

Jordan, N. (2017). *Google now provides AI-powered translations for Arabic and Hebrew*. VentureBeat.

Joshi, N., Mathur, I., Darbari, H., & Kumar, A. (2015). Incoprporating machine learning in machine translation evaluation. In E. El-Sayed, T. Sabu, T. Hideyki, P. Selwyn & H. Thomas (Eds.), *Advances in Intelligent Informatics*. Springer: UK. https://doi.org/10.1007/978-3-319-11218-3_20

Kafipour, R., & Jahanshahi, M. (2015). Error analysis of English translation of Islamic texts by Iranian translators. *Journal of Applied Linguistics and Language Research*, *2*(3), 238–252.

Koehn, P. (2009). *Statistical machine translation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

Llitj´os, A. F., Carbonell, J. G., & Lavie, A. (2005). *A Framework for Interactive and Automatic Refinement of Transfer-Based Machine Translation* (pp. 30–31). Proceedings of EAMT 10th Annual Conference.

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Traducció i qualitat Revista Tradumàtica: tecnologies de la traducció*, *12*. https://doi.org/10.5565/rev/tradumatica.77

Maučec, M., & Donaj, G. (2019). Machine translation and the evaluation of its quality. *IntechOpen*. Retrieved from https://www.intechopen.com/online-first/machine-translation-and-the-evaluation-of-its-quality

Munday, J. (2001). *Introducing Translation Studies: Theories and Applications*. Oxon: Routledge.

Nizke, J. (2019). *Problem solving activities in post-editing and translation from scratch: A multi-method study* (translation and multilingual natural language processing 12). Berlin: Language Science Press. https://doi.org/10.4324/9780429030376-5

Nord, C. (1991). *Text analysis in translation. Theory, methodology, and didactic application of a model for translation-oriented text analysis*. Translated from the German. Amsterdam/Atlanta: Rodopi.

Och, F. (2009). *51 Languages in Google Translate*. Google Research Blog on August 31. Retrieved from https://ai.googleblog.com/2009/08/51-languages-in-google-translate.html

Oudah, M., Almahairi, A., & Habash, N. (2019). The Impact of preprocessing on Arabic-English. Statistical and neural machine translation. *Proceedings of MT Summit XVII*, *1*, 19–23.

Popovic, M. (2018). Error Classification and Analysis for Machine Translation Quality Assessment. *Translation Quality Assessment*, 129–158. https://doi.org/10.1007/978-3-319-91241-7_7

Popovi´c, M., & Ney, H. (2006). *Error analysis of verb inflections in Spanish translation output* (pp. 99–103). TC-STAR Workshop on Speech-to-Speech Translation. Barcelona, Spain.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *94*. Retrieved from http://pareonline.net/getvn.asp?v=9&n=4

Turovsky, B. (2017). *Higher quality neural translations for a bunch more languages*. The Keyword Google Blog. Google. Retrieved March 6, 2019, from https://www.blog.google/products/translate/higher-quality-neural-translations-bunch-more-languages/

Turovsky, B. (2018). *Email interview: D'Monte, Leslie: Working to improve quality of translations*. Google's Barak Turovsky in EmTech India 2018—an emerging technology conference organized by Mint and MIT Technology Review. Retrieved October 10, 2019, from https://www.livemint.com/Companies/hrQAcRhDJm6acbKrwLKtlN/Working-to-improve-quality-of-translations-says-Googles-Ba.html

United Language Group. (2019). *Statistical vs. neural machine translation*. Retrieved August 3, 2019, from https://unitedlanguagegroup.com/blog/statistical-vs-neural-machine-translation/

Van der wees, M., Arianna, B., Wouter, W., & Christof, M. (2015). *What's in a domain? Analyzing genre and topic differences in statistical machine translation* (pp. 560–566). Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). https://doi.org/10.3115/v1/P15-2092

Vardaro, J., Schaeffer, M., & Hansen-Schirra, S. (2019). Translation quality and error recognition in professional

neural machine translation post-editing. *Informatics*, *6*(3), 41. https://doi.org/10.3390/informatics6030041

Vilar, D. et al. (2006). *Error analysis of statistical machine translation output* (pp. 697−702). LREC-2006: Fifth International Conference on Language Resources and Evaluation. Proceedings, (Genoa, Italy, May 2006). Retrieved August 27, 2013, from http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf

White, J. S., O'Connell, T., & O'Mara, F. (1994). *The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches* (pp. 193–205). Proceedings of the First Conference of the Association for Machine Translation.

Williams, M. (2009). *Translation quality assessment*. Ottawa: University of Ottawa Press.

Woehr, D., & Allen, H. (1994). Rater training for performance appraisal. *Journal of Occupational and Organizational Psychology*, *67*, 189–205. https://doi.org/10.1111/j.2044-8325.1994.tb00562.x

Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, Y. M., & Schuster, Z. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. Retrieved from https://arxiv.org/pdf/1609.08144.pdf

Zaghouani, W., Habashy, N., Obeid, O., Mohitz, B., Bouamor, H., & Oflazer, K. (2016). *Building an Arabic machine translation post-edited corpus: guidelines and annotation. European Language Resources Association* (ELRA) (pp. 1869–1876). Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). https://www.aclweb.org/anthology/L16-1295/.