Journal of Translation and Language Studies
jtls.sabapub.com
ISSN:2709-5681

J Trans & Lang Stud
(2021)2(1) : 19 - 32
doi:https://doi.org/10.48185/jtls.v2i1.122

Considering Machine Translation (MT) as an Aid or a Threat to the Human Translator: The Case of Google Translate

*Hamidreza Abdi
Freelance Researcher, Iran
Hooman_78h@yahoo.com

Received: 19.01.2020 • Accepted: 01.03.2021 • Published: 31.03.2021 • Final Version 31.03.2021

Abstract: The present study aims to evaluate the output quality of an online MT; namely, Google Translate, from English into Persian and compare its output with the translations made by the human translators to find out that whether MT applications are considered as an aid or a threat to human translators. For the application of the study, the researcher designed a translation test consisting of 60 statements from different types of texts proposed by Reiss (1989). The translation test was translated via Google Translate and administrated to three human translators to be rendered. The translations made by Google Translate and by the three human translators alongside the 60 statements were given to 40 judges to be evaluated based on Dorr et al. s' (2010) criterion of MT quality assessment, including semantic adequacy, fluency, and understandability. As results indicated, Google Translate gave a good overall performance in the translation of the 60 statements into Persian in terms of semantic adequacy and understandability, but not in terms of fluency. Thus, there should be no fear of human translators being replaced by the MT. As conclusion, MT applications cannot be considered a threat to human translators, but as an aid for them. The present study also offers some recommendations that can be beneficial to translation students, trainee translators, translation teachers, and professional translators.

Keywords*:* Computer-Assisted Translation (CAT) Tools; Machine Translation (MT); Google Translate.

1. Introduction

Using conventional methods, such as paper dictionaries and typewriters, cause translators not only to spend more time and energy but also produce costly translations. For example, translators have to use typewriters or record their own voices to be typed by typists later (Kay as cited in Abdi, 2019). Granell (2015) states that to produce high-quality translations in ever-shorter time, translators need to take into consideration quality and time requirements. In other words, traditional tools, as Kay infers, are in need of a fundamental change to be used in the translator's workstation.

The emergence of technology has brought about major changes in the translation industry the extent to which almost all traditional tools were replaced by modern ones. This has led to a shift in translators' attitude towards the use of new translation methods to produce high-

* Corresponding Author: *Hooman_78h@yahoo.com*

quality translations in a short time. The development of many types of translation tools was triggered by "the demand for a variety of translations by different groups of end-users" (Quah, 2006, p. 1). Machine Translation (MT) is "a sub-discipline of computational linguistics or, one could say, one of the latter's flagship application areas" (Nirenburg, 2003, p. 3).

The invention of MT caused great fears for most translators due to the assumption that the aim of creating MT, as Abdi (2019) states, was to replace the human translator. In this context, Granell (2015) implies that the use of MT instead of human translator has always been anticipated, and this anticipation will be continued. Hutchins and Somers (1992) argue that such fear was built on "a belief that computerization necessarily implies a loss of humanity" (p. 149). Hunt (2002) does not regard MT as a threat to the human translator because he believes that "computers will never replace translators, but translators who use computers will replace translators who don't" (p. 49).

For some time, access to MT was possible at a high cost for most translators. Thus, online MTs, such Google Translate, appeared to be freely available to translators. These online tools were welcomed by translators because of the opportunity they provided for them to save time and reduce translation costs. The important question that has always been raised is that whether they are able to produce high-quality translations. A high-quality translation is well defined by Koby et al. (2014, p. 416) as follows:

A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs.

As it is clear from the above definition, *accuracy* and *fluency* are two key paradigms that should be taken into consideration. Koby et al. (2014) argue that these two paradigms are kinds of quality that can evaluate and fulfill the purpose of the translation. Evaluating the output quality of MT is expensive and time consuming due to many aspects of translation, such as adequacy, fidelity, fluency, and need to be evaluated (Hovy, 1999). Despite different types of automatic evaluation metrics, human evaluation, as Han (2018) discusses, is "usually trusted as the golden standards." Popovic et al. (2013) state that human translators are "the key to evaluating MT quality and also to addressing the so far unanswered question when and how to use MT in professional translation workflows" (p. 231).

1.1. The Present Study

Dorr et al. (2010) consider three important paradigms for evaluating the output quality of MT, including semantic adequacy, fluency, and understandability, of which semantic adequacy is the most important paradigm that "is widely regarded as the Gold Standard for assessment of MT quality" (p. 811). In the light of these criteria of MT quality assessment, the aim of the present study was to evaluate the output quality of an online MT, in this case Google Translate, from English into Persian. This has led to find out that whether Google Translate is considered as an aid or a threat to human translators. To achieve the objective of the present study, the researcher attempted to answer the following questions:

1. What was the output quality of Google Translate based on the three paradigms proposed by Dorr et al. (2010), including semantic adequacy, fluency, and understandability, for assessment of MT quality?
2. What can Google Translate be considered for the human translators, an aid or a threat?

The findings of the present study should help translation students and trainee translators to improve their knowledge of using MTs in their translations and turn the translation teachers' attention back to translation practices to design new materials, including MTs and new technologies. Furthermore, findings encourage professional translators to make contribution to the improvement of Google Translate function.

2. Review of Literature

2.1. Google Translate

Regardless of what translation quality Google Translate produces, it covers more than 100 languages today. Along with translating texts, Google Translate provides its users with several facilities, such as pronouncing the translated texts, highlighting the corresponding words in the source text (ST) and target text (TT), automatic language detection, suggesting further translations, voice recognition translation, and image translation (Al Mahasees, 2020; MAHDY, Samad, & Mahdi, 2020). He further states that Google Chrome and Mozilla are two most popular web facilities that give users the opportunity to use Google Translate. One advantage of Google is that it has built "an active translation community and suggested translation service" (Al Mahasees, 2020, p. 19) that enables users to choose up to five languages to offer accurate translations.

According to Grajales (2015), Google Translate is "a well-known app from Google, which works as a multi-language functional translator." Lotz and Van Rensburg (2014) imply that Google Translate is "a free online application, offered by Google Inc. that allows users to have words, sentences, documents and even websites translated in an instant"(p. 237). They describe that Google Translate encompasses computer systems that produce translations based on patterns including large amounts of text, not on a series of rules for a particular language. In other words, a large amount of texts has been stored in the storage of Google Translate to be retrieved by translators when they are in need.

Wu et al, (2016) argue that Google Translate has changed its approach from the Statistic Based Machine Translation (SBMT) to the Neural Machine Translation (NMT) to bridge the gap between human and machine translation and deal with the problems MT faces, such as rare words. In this context, Franz Josef Och (cited in Helft, 2010), a German computer scientist, implies that this technology can remove the language barrier, and it would provide the opportunity for anyone to communicate with anyone else.

2.2. Human Judgments for Manual Assessment of MT Quality

Human judgments are effective methods for developing the function of MT. In this regard, Coughlin (2001) mentions that "human evaluation has been the only means of providing the necessary feedback to keep development moving forward" (p. 63). Graham et al. (2013) argue that human evaluation for manual assessment of MT quality is "a key element in the development of machine translation systems" (p. 16). Evaluation of MT involving human judgments is "an extremely demanding task" (Brkic et al., 2013, p. 1). Furthermore, such

evaluation, as Dorr et al. (2010) argue, causes two main concerns that must be taken into consideration.

The first concern is that different answers, as Dorr et al. (2010) discuss, may be given by humans and different opinions may be expressed by them for the same evaluation. Thus, a panel of independent judges is needed to calculate the average of those differences. The second concern, as Dorr et al. imply, is related to the familiarity of the judges with the subject matter and/or sub-languages that can affect evaluation. In spite of time-consuming and costly as well as such concerns about human judgments, human evaluation is "a gold standard for evaluation of MT quality" (Brkic et al., 2013, p. 1) and used as "a baseline by which evaluation metrics are frequently judged"; therefore, there is no other evaluation metric to replace human judgments of translation (Dorr et al., 2010, p. 808).

2.3. Dorr et al. s' Criterion of MT Quality Assessment

To assess the output quality of MT, many paradigms need to be observed, such as semantic adequacy, fluency, and understandability (Dorr et al., 2010). According to Callison-Burch (2007), the first two are the most common paradigms that are used in human evaluation method.

The most important paradigm, as Dorr et al. (2010) imply, is semantic adequacy (or fidelity) which seeks the answer to the question: does translation have the same *meaning* as the source-language material? Semantic adequacy evaluation of the MT output has nothing to do with fluency judgment and ignores it because adequacy evaluation "measures whether the essential information in the source can be extracted from the system output" (p. 808). Moreover, evaluating semantic adequacy, as they state, is more challenging than evaluating accuracy due to that the evaluator needs to be bilingual in order to judge whether information is correctly transferred from the source language (SL) to the target language (TL).

Fluency is also considered an important paradigm, so that translation should be something that native speaker of the TL would say or write alongside its understandability (Dorr et al., 2010). They discuss that the SL input does not require judge fluency, so that "ratings for fluency are usually obtained from monolingual judges" (p. 808). Thus, the evaluator, as they argue, needs to be a fluent speaker in TL to judge whether translation is fluent, without taking into account the accuracy of the translation. In order to judge fluency and adequacy, a five- or seven-point scale should be prepared, and they should also be measured separately on each sentence in the system output (Przybocki cited in Dorr et al., 2010). In this context, Dorr et al. (2010) imply that typicall the human judgments are provided with a multi-point scale, but they are sometimes asked to make their judgments on some sort of numeric scale that leads to correlation/regression analysis.

Understandability is another paradigm that should be taken into consideration. Meaning of translation may be completely and correctly conveyed but may be so awkward and difficult to understand (Dorr et al., 2010). For example, a literal translation may be used for a SL idiom that makes its meaning understandable to target readers. To judge understandability, it is enough to receive a yes/no response from judges according to Dorr et al. (2010).

To judge the above paradigms, judges can be monolingual or bilingual. If they are monolingual, they need to be given one or more high-quality reference human translation; and if they are bilingual, they should be provided with the SL material.

3. Methodology

3.1. Participants

The participants of the present study were 40 judges who were invited to evaluate the output quality of Google Translate on the basis of a five-point scale prepared to examine the three paradigms proposed by Dorr et al. (2010). As the official website of Iranian Association of Certified Translators and Interpreters (IACTI) was available on the Internet (https://www.iacti.ir), the researcher decided to choose the judges from certified translators. The logic behind this selection was that they were qualified by the competent authorities in advance that facilitated the process of selection of the judges/participants for the present study. Moreover, such certified translators helped the researcher fulfil the objectives of the present study and achieve the intended results. Before they judged, they were informed about the subject matter sufficiently to obtain valid judgments.

3.1. Instruments

A translation test consisting of 60 statements was prepared for data collection. The statements were extracted from different types of texts proposed by Reiss (1989). Thus, all statements were divided into three categories: informative, expressive, and operative texts. Each category was organized into four sub-categories: simple, complex, compound, and complex-compound statements. In other words, five statements of each type were chosen for each category. This led to examine the function of Google Translate in the translation of not only different text types but also types of statements and find out the weaknesses and strengths of Google Translate. All statements were extracted from online websites, such as news websites and online dictionaries. To validate the translation test, a panel of university professors, who had teaching experience in Translation Studies, was asked to determine the content validity of the statements. The feedback received from translation teachers was constructive and helped to improve the translation test in terms of rewording and reordering of the statements. The test-retest method was applied to measure the reliability of the questionnaire. It was given to 10 certified human translators. After two weeks, the test was administrated to the same translators. The results of the two trials were correlated and the coefficient of correlation indicated the reliability of the test (r = .802).

3.4. Procedure

The following steps were established for data collection: first, the translation test was translated via Google Translate. The translations alongside the 60 statements were given to the judges to evaluate the output quality of Google Translation based on Dorr et al. s' (2010) criteria of MT quality assessment. Thus, a five-point scale on a continuum from "None" to "All" for evaluating semantic adequacy was prepared that showed how much the meaning was correctly expressed in the translation. A five-point scale on a continuum "Incomprehensible" to Flawless Persian" was also designed to find out the extent to which the translation was fluent. The understandability of the output of Google Translate was made based on the "Yes/No" judgment. To analyze the data of the present study, the frequencies and percentages of each continuum chosen by the judges were calculated and illustrated in tabulation forms. Inferential statistics, such as one sample Wilcoxon signed ranks test, were employed to justify the hypothesis.

4. Results

4.1. Semantic adequacy

As table 1 indicates, the judges agreed with the semantic adequacy of a large number of informative statements (84%) (42% were completely adequate, 22% most adequate, and 20% much adequate); whereas a few number of them (16%) were inadequate from the judges' points of view. Based on Table 1, more than half of the expressive Persian translations (65%) were chosen by the judges as adequate ones (23% were completely adequate, 21% most adequate, and 21% much adequate). The rest of the statements (35%) were inadequate according to the judges' opinions. Table 1 also illustrates the agreement of the judges with the equality of the number of completely inadequate (29%) and adequate (29%) operative Persian translations made by Google translate. Furthermore, the third answer (Much) was chosen for a few numbers of the Persian translations (5%); followed by the second answer (Little) (15%).

Table 1. Answer Frequencies of the Judges to Semantic Adequacy of the 60 Statements

Statement Types	*N*	None		Little		Much		Most		All		*N**		
		f	*%*	*f*	*%*	*f*	*%*	*f*	*%*	*f*	*%*	*f*	*%*	*M*
Informative	20	45	60.0	805	10.0	158	20.0	178	22.0	339	42.0	800	100.0	3.85
Expressive	20	153	19.0	131	16.0	165	21.0	164	21.0	187	23.0	800	100.0	3.12
Operative	20	231	29.0	117	15.0	40	5.0	183	23.0	229	29.0	800	100.0	3.07
Total	60	429	18.0	328	14.0	363	15.0	525	22.0	755	31.0	2400	100.0	3.34

Note. N=number of expressive sentences; N*= number of total answers.

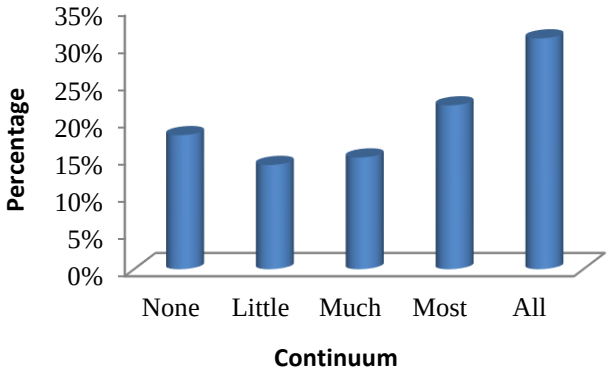

Figure 1. Percentages of the Judges' Evaluation of Semantic Adequacy of the 60 Statements

One sample Wilcoxon signed ranks test was run to test the hypothesis. In other words, the test allowed the researcher to see whether there was a significant relationship between the judges' agreements and the semantic adequacy of Persian translations of each statement type produced by Google translate. Table 2 presents the results of one sample Wilcoxon signed ranks test, which shows that the hypothesis was retained in the case of expressive and operative statements because the *p* values were higher than .05 ($p > .05$). This,

alongside the mean scores of the expressive and operative statements, illustrates that the judges' agreement with the Persian translations of these two types is not significant, indicating average semantic adequacy (*ME* = 3.12; *MO* = 3.07). The mean score of informative statements was higher than the mid-answers (i.e. 2.5) to the sentences, considering the theoretical mean/median of the population. The hypothesis was also rejected due to the value of *p* that was lower than .05 ($p < .05$). This implies that the judges' agreement with semantic adequacy of the informative statements is significant and Google translate mostly produced adequate informative translations (*MI* = 3.85).

Table 2. One Sample Wilcoxon Signed Ranks Test for Semantic Adequacy of Each Statement Type

Statement Types	N	MDN	p
Informative	20	4.5	.000
Expressive	20	3	.07
Operative	20	4	.08

Note. N= total number of statements; MDN= median; The sig value of one sample Wilcoxon signed ranks test is significant at $p < .05$.

In general, as Table 1 indicates, the judges were in full agreement about the semantic adequacy of almost two-thirds of the 60 statements (68%); whereas they disagreed with semantic adequacy of 31% of the statements. One sample Wilcoxon signed ranks test was also employed to test the hypothesis, whether the agreement of the judges with the degree of semantic adequacy of the 60 Persian translations that Google translate produced was significant. To do so, the sign value of this test was calculated. According to Table 3, the *p* value of the test was .0 that was lower than .05 ($p < .05$). Thus, the hypothesis was rejected. Furthermore, the mean score of total answers to the 60 statements was 3.34 out of the highest value 5 that was higher than the mid-answers to the sentences (3.34 > 2.5) (see Table 1). In other words, the judges in general significantly agreed with the semantic adequacy of the Persian translation of the 60 statements made by Google translate (*MT* = 3.34).

Table 3. Total Score of One Sample Wilcoxon Signed Ranks Test (Semantic Adequacy)

N	*MDN*	*p*
60	4	.000

Note. N= total number of statements; MDN= median; The sig value of one sample Wilcoxon signed ranks test is significant at $p < .05$.

4.2. **Fluency**

The judges did not agree with fluency of 61% of the informative statements. That is, 31% of the statements was considered incomprehensible and disfluent, and 30% non-native English. For the judges, only 39% of the sentences was fluent (4% was flawless Persian and 35% good Persian) (see Table 4). According to answer distribution, all Persian translations of expressive statements put into the following continuums: incomprehensive (37%), disfluent Persian (24%), non-native Persian (21%), good Persian (15%), and flawless Persian (3%). According to the judges' opinion, a great majority of the operative translations (73%) was incomprehensible (37%), disfluent (26%), and non-native Persian (10%). By contrast, a small number of operative Persian translations (27%) was produced in a fluent manner (9% was flawless Persian and 18% good Persian). Table 4 demonstrates the

agreement of the judges with the disfluency of two-third of the 60 Persian translations (28% was incomprehensible, 24% disfluent, and 20% non-native Persian). The judges gave scores to 22% as good Persian, and to 5% as flawless Persian.

Table 4. Answer Frequencies of the Judges to Fluency of the 60 Statements

Statement Types	*N*	Incomprehensible		Disfluent Persian		Non-native Persian		Good Persian		Flawless Persian		*N**		
		f	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%	*M*
Informative	20	82	10.0	165	21.0	238	30.0	283	35.0	32	4.0	800	100.0	3.02
Expressive	20	298	37.0	191	24.0	170	21.0	117	15.0	24	3.0	800	100.0	2.22
Operative	20	292	37.0	211	26.0	83	10.0	141	18.0	73	9.0	800	100.0	2.36
Total	60	672	28.0	567	24.0	491	20.0	541	22.0	129	5.0	2400	100.0	2.53

Note. N=number of statements; N*= number of total answers.

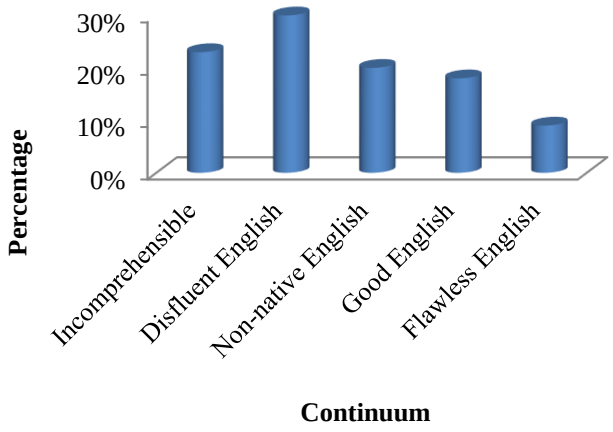

Figure 2. Percentages of the Judges' Evaluation of Fluency of the 60 Statements

Table 5. One Sample Wilcoxon Signed Ranks Test for Fluency of Each Statement Type

Statement Types	*N*	*MDN*	*p*
Informative	20	3	.04
Expressive	20	2	.07
Operative	20	2	1.0

Note. N= total number of statements; MDN= median; the sig value of one sample Wilcoxon signed ranks test is significant at $p < .05$.

One sample Wilcoxson test was run to test the hypothesis in order to see the extent to which the judges expressed their agreements with the fluency of each statement type and whether there was a significant relationship between the judges' agreements and the fluency of Persian translations of each statement type. The results of the test indicate that the *p* values relating to expressive and operative Persian translations were .07 and 1.0 that seems to be

higher than .05 ($p > .05$) (see Table 5). Thus, the hypothesis was not rejected. This points to that there is no significant relationship between the judges' opinions and the fluency of Persian translations of these two types of statements. Moreover, the mean scores of given answers to these translations were lower than theoretical mean/median (i.e. 2.5). This shows disfluency of the expressive and incomprehensibility of operative translations that Google translate produced (*ME* = .07, *MO* = 1.0). In contrast, the results show the amount of *p* value of informative statements lower than .05 ($p < .05$). Hence, the hypothesis was rejected, and the judges' opinion was significantly different from the theoretical mean/median. According to Table 4, the mean score of this sentence type was higher than the theoretical mean/median (i.e. 2.5) that gave the indication of average fluency of the Persian translations made by Google translate (*MI* = 3.02).

Table 6. Total Score of One Sample Wilcoxon Signed Ranks Test (Fluency)

	N	*MDN*	*p*
Statements	60	3	.35

Note. N= total number of statements; MDN= median; The sig value of one sample Wilcoxon signed ranks test is significant at $p < .05$.

Considering the result of total answers to the 60 statements, one sample Wilcoxson test was also applied to test the hypothesis. That is, there was a significant relationship between the judges' agreements and the fluency of the 60 Persian statements made by Google translate. Table 6 indicates that the *p* value was higher than .05 ($p > .05$). As a result, the hypothesis was retained. In addition, the total mean score of the given answers was 2.53, almost equal to the theoretical mean/median (i.e. 2.5). Thus, the judges' agreement was not significant and not differs from the theoretical mean. This indicates an average fluency of the Persian translations that Google translate produced (*MT* = 2.53).

4.3. **Understandability**

As Table 7 shows, more than half of the statements (55%) was understandable; whereas 44% of them was not easy to understand from the judges' viewpoints.

Table 7. Answer Frequencies of the Judges to Understandability of the 60 Statements

	N	Yes		No		*N**		
		f	*%*	*f*	*%*	*f*	*%*	*M*
Total	60	1322	55.0	1078	45.0	2400	100.0	1.44

Note. N=number of expressive sentences; N*= number of total answers.

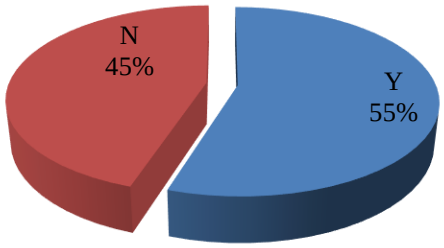

Figure 2. Percentages of the Judges' Evaluation of Understandability of the 60 Statements

Table 8 illustrates the results of one sample Wilcoxon test. The test was used to test the hypothesis, whether a relationship between the judges' agreements and the understandability of the 60 Persian statements was significant. Based on Table below, the *p* value of the test was higher than .05 ($p > .05$). Hence, the hypothesis was not rejected, and the judges' opinion was not significantly different from the theoretical mean/median (i.e. 1.5). The mean score of the given answers was lower than (almost equal to) the theoretical mean/median, showing average understandability ($MT = 1.44$).

Table 8. One Sample Wilcoxon Signed Ranks Test (Understandability)

	N	*MDN*	*p*
Statements	60	1	. 51

4. Discussion

In the previous section, the output quality of Google Translate was analyzed in terms of semantic adequacy, fluency, and understandability. This section discusses the results derived from the analysis and reports the findings of some recent studies in the field.

According to the results, Google Translate had significant performance in the translation of almost two-thirds of the 60 statements into Persian. In other words, the highest scores given to the output of Google Translate were *Almos*t and *All* (see Examples 1 and 2). It means that the meaning of the ST is completely conveyed to the TT. In spite of a few number of translations that were given the lowest scores, the fragments of the ST meaning were appeared in the TT (see Example 3). Zakaryia (2020) reports the same findings in his doctoral dissertation. The findings derived from the comparison between the output quality of the three MTs, one of which was Google Translate, in 2016 and 2017. According to the results, the score for Google translate was *everything score* in both years and the performance of Google Translate was better in 1970 than 1960.

Example 1

"Iran released a dozen prominent political prisoners last week." (Informative, simple)

"ایران هفته گذشته دوازده زندانی سیاسی برجسته را آزاد کرد." (All)

Example 2

"Evergreen trees are a symbol of fertility because they do not die in the winter" (Expressive, complex)

"درختان همیشه سبز نماد باروری هستند زیرا در زمستان نمی میرند." (Almost)

Example 3

"Smiling is keeping one moment, but its happiness spreads about whole day." (Operative, compound)

"لبخند زدن یک لحظه را حفظ می کند ، اما شادی آن در کل روز گسترش می یابد." (Little)

As the results show, Google Translate produced a minority of translations fluently in such a way that the translations were exceedingly erroneous in terms of fluency (see Examples 4 and 5). Furthermore, there was a low degree of fluency involved in the translations (Example 6). This gave the indication of the lack of fluency in the translations made by Google Translate. In this context, Kadhim et al. (2013) conduct a study on two types of MTs, Google Translate and Babylon, to find out the differences in their performance by making a comparison between the output qualities of the two MTs. The results obtained from their study about the degree of fluency were similar to the results of the present study. That is, Google Translate produced the low-quality translations in terms of fluency the extent to which the average of Google's fluency was the lowest average among the other paradigms, such as semantic adequacy and understandability.

Example 4

What should be the punishment of such a person?" (Expressive, simple)

"مجازات چنین شخصی چه مجازاتی باید داشته باشد؟" (Disfluent Persian)

Example 5

"You may choose to fear half empty or celebrate half full, but either way you get to choose." (Operative, compound)

"ممکن است شما ترجیح دهید نیمه خالی بترسید یا نیمه کامل جشن بگیرید ، اما به هر روشی که انتخاب کنید." (Incomprehensible)

Example 6

This is the same evil hand that singed economic punishments against Iranian nations." (Informative, complex)

"این همان دست شیطانی است که مجازات های اقتصادی علیه ملت ایران را امضا کرد." (Flawless Persian)

Understandability was the last paradigm that the judges were asked to evaluate. Dorr et al. (2010) imply that the quality of semantic adequacy is the most important paradigm that needs to be judged, but this quality itself "is measured in an interestingly indirect way that encompasses the understandability of the translation" (p. 814). It is rather to say, there is an indirect relation between semantic adequacy evaluation and understandability evaluation. Based on the results, the translations made by Google Translate were informative and easy to understand to some degree (see Example 7). This indicates not only the average clarity of the Google Translate output but also the quality of semantic adequacy of the Persian translations.

Example 7

"Government does not solve problems; it subsidizes people." (Informative,

compound)

"دولت مشکلی را حل نمی کند ؛ بلکه به مردم یارانه می دهد." (Understandable)

The main shortage of Google Translate was in the translation of statements that contained implied meanings, such as operative texts and idioms. In such cases, the translations made by Google Translate were in need of post-editing because neither their meanings were fully conveyed nor they were fluent and comprehensible. Furthermore, Google Translate is a rule-based MT and works based on some defined rules. This affects the output quality of such a popular online MT.

5. Conclusion

Notwithstanding automatic evaluation of the output quality of MT that is considered a cost-effective method, human evaluation has some advantages one of which is the capability of the human evaluators to "perform some tasks that are currently beyond the reach of automated metrics," such as monitoring the quality of translated individual sentences with high accuracy (Coughlin, 2013, p. 69). Hence, the present study attempted to evaluate the output quality of Google Translate to see whether MTs are considered an aid or a threat to human translators. As the results indicate, translations Google Translate produced were acceptable to a certain degree. In a wider sense, the performance of Google Translate was partly satisfactory but not the extent to which it is used alone. However, the output of Google Translate needs to be edited by human translator to reach a high-quality output. To approve this, the researcher relies on the report from the Automatic Language Processing Advisory Committee (1966) where it is clearly mentioned that "it was not possible to obtain a translation that was entirely carried out by a computer and of human quality" (cited in Delpech, 2014, p. 4). Post-editing, as Hutchins (1996) implies, is the only solution to reach a good quality of translation. In conclusion, the assumption that was made on the replacement of MTs with human translators is rejected and MTs are considered an aid not a threat to human translators.

The present study offers some pedagogical implications that can be helpful to translation students, trainee translators, translation teachers, and professional translators.

To translation students and trainee translators, it is recommended that they improve the skill of using not only MTs but also other technological tools appropriately. They are not recommended to rely on Google Translate alone to produce high-quality translations. To reach good-quality translations, they need to apply their own skills alongside Google Translate. The latter helps to reduce costs and time of the translation while the former provides the translator with the opportunity to enhance translation quality.

Teaching new technologies is the task that translation teachers are responsible for. They need to take serious attention to translation practices alongside translation theories. To do this, translation teachers can encourage students to attend seminars and workshops held by experts in the field to be familiar with new technologies and broaden the knowledge of using such tools. By designing new materials, encompassing modern technological tools as well as giving translation tasks that involve students with Google Translate and other types of MTs, translation teachers can teach students to employ MTs in their translation instead of using traditional methods.

The role the professional translators play is of great importance to the improvement of Google Translate function. Thus, they are kindly recommended to make contribution to enhance the output quality of this popular MT by validating translations. That is, professional translators select the correct translation from those offered by Google Translate for the individual statement, especially those encompassing implied meanings, such as proverbs and idioms. This helps not only improve the function of Google Translate but also make it possible for people around the world understand language a little better.

Authors' contributions: Not applicable

Conflicts of Interest: The authors declare no conflict of interest.

The funding sponsors: The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

[1] Abdi, H. (2019). Translation and technology: A study of Iranian freelance translators. Mauritius: Lambert Academic Publishing.

[2] Al Mahasees, Z. (2020). Diachronic evaluation of Google Translate, Microsoft Translator, and Sakhr in English-Arabic translation. Unpublished Master's Thesis, the University of Western Australia, Australia.

[3] Brkic, M., Seljan, S., & Vicic, T. (2013, March 24-30). Automatic and human evaluation on English-Croatian legislative test set [Conference session]. 14th International Conference, Samsos, Greece. https://www.bib.irb.hr/547340.

[4] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder. J. (2007). (meta-) evaluation of machine translation. Proceedings of the Second Workshop on Statistical Machine Translation (pp. 136–158). Prague, Czech Republic.

[5] Coughlin, D. (2001). Correlating automated and human assessments of machine translation quality. Proceedings of MT Summit IX (pp. 63–70). New Orleans, LA.

[6] Delpech, E.M. (2014). Comparable corpora and computer-assisted translation. London: ISTE.

- Dorr, B., Snover, M., & Madnani, N. (2010). Machine translation evaluation. In J. McCary, C. Christianos & J. P. Olive (Eds.), Handbook of Natural Language Processing and Machine Translation (pp. 801-814). Heidelberg: Springer.

[8] Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2013). Crowd-sourcing of human judgments of machine translation fluency. Proceedings of Australasian Language Technology Association Workshop (pp. 16-24). Brisbane, Australia.

[9] Grajales, G. C. A. (2015, Jun 23). The statistics behind Google Translate. Retrieved May 3, 2020, from http://statisticsviews.com/details/feature/8065581/The-statistics-behind-Google Translate.

[10] Granell, X. (2015). Multilingual information management: Information, technology and translators. Kidlington, OX: Chandos Publishing.

[11] Han, L. (2018, September 19). Machine translation evaluation resources and methods: A survey [Conference session]. Ireland Postgraduate Research Conference (IPRC), Dublin, Ireland. https://www.researchgate.net/publication/ 303280649 Machine Translation Evaluation A Survey.

[12] Helft, M. (2010, March 9). Google's computing power refines translation tool. The New York Times. http://www.nytimes. com/2010/03/09/ technology / 09 translate.html

[13] Hovy, E. H. (1999). Toward finely differentiated evaluation metrics for machine translation. Proceedings of the Eagles Workshop on Standards and Evaluation. Pisa, Italy.

[14] Hunt, T. (2002). Translation support software: the next generation replacement to CAT tools. ATA Chronicle, 31(1), 49–52.

[15] Hutchins, J. (1996). ALPAC: the (in)famous report. MT News International, (14), 9–12.

[16] Hutchins, J., & Somers, H. (1992). An introduction to machine translation. London: Academic Press.
[17] Kadhim, K. A., Habeeb, L. S., Sapar, A. A., Hussin, H., & Abdullah, M. R. T. (2013). An evaluation of online machine translation of Arabic into English news headlines: Implications on students' learning purpose. The Turkish Online Journal of Educational Technology, 2(12), 39-50.
[18] Koby, G. S., Fields, P., Hague, D., Lommel, A., & Melby, A. (2014). Defining translation quality. Revista Tradumàtica: tecnologies de la traducció 12, 413-420.
[19] Lotz, S., & Van Rensburg, A. (2014). Translation technology explored: Has a three-year maturation period done Google Translate any good? Stellenbosch Papers in Linguistics Plus, 43, 235-259.
[20] MAHDY, O. S. M. M. S., Samad, S. S., & Mahdi, H. S. (2020). The attitudes of professional translators and translation students towards computer-assisted translation tools in Yemen. *Dil ve Dilbilimi Çalışmaları Dergisi, 16*(2), 1084-1095.
[21] Nirenburg, S. (2003). Introduction. In S. Nirenburg, H. Somers & Y. Wilks, (Eds.), Readings in machine translation (pp. 3-11). Massachusetts: Massachusetts Institute of Technology.
[22] Popovic, M., Avramidis, E., Burchardt, A., & Hunsicker, S. (2013). Learning from human judgments of machine translation output. Proceedings of the MT Summit XIV (pp. 231–238). Nice, France.
[23] Quah, C.K. (2006). Translation and technology. New York: Palgrave MacMillan.
[24] Reiss, K. (1989). Text types, translation types and translation assessment. In A. Chesterman (Ed.), Readings in translation theory (pp. 105–115). Helsinki: Finn Lectura.
[25] Wu, Y., Schuster, M., Chen, Z., V. Le, Q., & Norouzi, M. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144, 1-23.
[26] Zakaryia, A. (2020). Diachronic evaluation of Google Translate, Microsoft Translator, and Sakhr in English-Arabic translation. Unpublished PhD Dissertation, the university of Western Australia, Western Australia, Australia.

About the Author

Hamidreza Abdi holds a master's degree in Translation Studies from Islamic Azad University, Science and Research, Tehran, Iran. He received his BA in the same major from Islamic Azad University, Roodehen Branch, Iran in 2009. He is a freelance researcher in the field of translation studies. He has also published numerous articles in different areas of translation. His main research interest is translation and technology..