A Synchronic Assessment of Google Translate, Microsoft Translator and Reverso in English<>Arabic Translation

Zakaryia Almahasees[1]

[1]Department of English Language and Translation, Applied Science Private University, Amman, Jordan

***Abstract*— The study aimed to evaluate three multilingual MT systems that provide free English<> Arabic translation services. It assessed the effectiveness of the three systems in dealing with WHO, UN, Petra News Agency, and Arab League documents. The study adopted (LDC, 2002) intelligibility and fidelity scales to highlight the limitations and strengths of the systems understudy. The study found that Google Translate outperformed the three systems in providing intelligible and fluent output without flaws with 230 segments, equaling 57% of the corpus size, followed by Microsoft Translator with 48% and Reverso with 45% translation free of errors. On the other hand, the study found that Google Translate outperformed the other systems in terms of adequacy and fluency scales, followed by Microsoft Translator and Reverso interchangeably. However, the three systems are still unable to substitute human translators. Therefore, the study recommends conducting a diachronic study to trace the three systems' development over five years scientifically.**

***Keywords*— Error Analysis, Fidelity, Intelligibility, Machine Translation,**

I. INTRODUCTION

Knowledge exchange and communication across human societies are established using translation, an effective medium. Technology has had a pervasive impact on human life during the 20th century(Z. Almahasees & Jaccomard, 2020). Technological advancements have increased the desire to use technology for human services. Many entities, including people, businesses, professions, and governments worldwide, use the Internet as one of the most cutting-edge inventions. Because the Internet is widespread in all aspects of our lives, people today live in a globalized society, occasionally called a "village." Importantly, people can obtain a wide range of information provided they can access the Internet(Z. M. Almahasees, 2017, 2018).

The translation is one of the domains affected by the spread of technology. Integrating technology in translation dates to the first half of the 20th century. Thanks to the Internet and MT services, the fruit of this integration is that Humans can read documents written in other languages in their languages. Moreover, there are numerous translation systems, some of which require paid subscriptions, others offer free access, and still, others are desktop translation programs, which somewhat restrict their popularity, functionality, and use. Because online Machine Translation (MT) systems can be accessed for free or at a minimal cost, it is essential to assess the capacity of MT systems to provide adequate translation to the end users(Almahasees, 2021).

Machine Translation (MT) uses a computer to render human languages. MT, a type of natural language processing (NLP) tool, uses computers to provide translation for human languages automatically. In machine translation (MT), dictionaries that are either monolingual or bilingual, corpus-based, neural networks, and other algorithm-based procedures are used to offer translation from one language to another. It involves applying linguistics rules using the Rule-Based Machine Translation (RBMT) Approach rather than merely changing one word for another. Additionally, it uses statistical models through the Statistical Based Machine Translation (SBMT) method to generate translations based on already established or electronically accessible corpora. When translating from a source language (SL) into a target language (TL), MT uses linguistic knowledge of syntax, morphology, and semantics (RBMT technique) and Statistical models (SBMT approach) (Almahasees, 2021; Dabre, Chu, & Kunchukuttan, 2020). Currently, MT systems use Neural Machine Translation (NMT) which uses artificial networks to produce equivalent translations like human translation (Altintas & Cicekli, 2022).

The development of MT can be traced back to the media revolution, the growth of NLP applications, the advancement of artificial intelligence, and the research requirement to provide cross-cultural communication. Because of this, MT systems have developed quickly and multiplied. The three widely utilized systems (Greatcontent, 2022) selected for the current study are Google, Microsoft Translator, and Reverso. Each system offers automated translation for both English and Arabic on various platforms, including desktop, mobile devices, online, speech translation, voice, and offline. The study aims to assess the output of two central MT systems: Google Translate, Microsoft Translator, and Reverso.

Neural Machine Translation is a method of machine translation that models complete sentences in a single integrated model and uses artificial neural networks to estimate the likelihood of word sequences. Google Translate is a multilingual platform providing free translation services for 133 languages. It has 500 million daily users who can translate billions of words and phrases daily (GoogleTranslate, 2022). Microsoft Translator is a multilingual platform that facilitates the translation of over 103 languages (MicrosoftTranslator, 2022). Reverso is also a multilingual platform that provides translation for 19 languages. The number of monthly users reaches 48 million

DOI: http://doi.org/10.24086/ICLANGEDU2023/paper.931

(Reverso, 2022). The chosen systems utilize NMT, which uses neural networks to identify the equivalent terms across human languages. The study evaluates the development of the three systems synchronically in terms of adequacy and fluency scales over 2022.

Throughout MTE history, numerous assessment techniques for machine translation have been used. Each translation system's effectiveness, acceptance, and cost-efficiency are essential to MTE. The evaluation of MT systems extended over 70 years since the early attempts of MT in 1949. Manual evaluation entails human intervention to assess MT output, while automatic evaluation relies on using automated metrics to measure the closeness of MT output to the source text (Z. Almahasees, Husienat, & Husienat, 2022; Chatzikoumi, 2020). Automatic evaluation is cheap, time-saving, and objective, while human evaluation is costly, time-consuming, and inconsistent. However, the automated assessment does not reflect the quality of the translation. Instead, it reflects text similarity between the output and the source texts in terms of text similarity (Sin-Wai, 2015). Therefore, the study adopts manual evaluation to ensure the best translation service the three systems provide.

Manual evaluation entails human evaluators to assess the MT output(Z. Almahasees, 2021). Manual evaluation can be conducted in various ways: Fidelity, intelligibility, translation ranking, error analysis, and post-editing. The Fidelity evaluates the language quality of the output. It aims to ensure that the output is free of any grammatical flaws or errors that inhibit the fluency of the text. The intelligibility part assesses the transference of meaning in comparing the ST with the MT output. Finally, the translation ranking works to identify the best sentences from the worst in terms of grammatical construction, which is like error analysis. The study uses the intelligibility and fidelity scales to conduct a synchronic evaluation for the three chosen systems over three years.

Intelligibility measures the degree of acceptability of the translation to the end users(Z. Almahasees & Jaccomard, 2020). Intelligibility scales measure the transfer of meaning based on a 4-point scale. The intelligibility scales were flawless, good, non-native, disfluent, and incomprehensible. The flawless scale indicates that the translation is free of any grammatical errors; good English (minimum number of errors which does not inhibit the fluency of the text); non-native (the translation indicates that the text is not fluent); disfluent (some fragments of ST contained in the TT); incomprehensible (challenging to understand due to the wrong formation of the translation in terms of grammatical rules). On the other hand, Fidelity evaluates the transfer of the meaning without any additions, omissions, or meaning loss. Fidelity has a 5-point scale: all, most, much, little, and none. ***All*** scale refers to the transfer of the meaning contained in the ST.

Scale	Fidelity	Intelligibility
5	All	Flawless
4	Most	Good
3	Much	Non-native
2	Little	Disfluent
1	None	Incomprehensible

Figure 1. Fidelity and Intelligibility Scale (LDC, 2002)

II. Literature

Arabic has been studied since the beginning of MT. However, Arabic Machine Translation (AMT) needs much collaborative research to reach adequate levels and void the gap with other languages. The growth of the AMT knowledge base is hindered by the tiny number of surveys and research that individual researchers in the field conduct. The high demand is for joint research between individual researchers and academic institutions to enhance AMT. Several studies were conducted to assess the output of MT synchronically. However, there is a lack of synchronic studies on MT.

Salem, Hensman, and Nolan (2008) addressed the challenges of Arabic language restrictions and how they affect MT's growth from Arabic to English. They established a model including the role and reference grammar and ways to address troublesome difficulties with Arabic and offered remedies for the MT's misunderstanding over Arabic word order. Adly and Al Ansary (2009) conducted their study evaluating Arabic-English MT based on the Interlingua method of machine translation and the Universal Networking Language (UNL). For various lines from the Encyclopedia of Life Support Systems, they compared the MT outputs of three translation systems: Google, Tarjim, and Babylon (EOLSS). The three widely used automatic evaluation measures, BLEU, F1 measure, and F Measure mean, were modified by researchers to account for the syntactic and semantic limitations of Arabic. They discovered that the current systems do not translate the semantic cohesiveness and style of Arabic well.

Farghaly and Shaalan (2009)assert that due to the unique characteristics of the Arabic language, NLP methods created by Western corporations are not "easily applicable to Arabic." Al-Alamiyyah Group, as was already noted, was the first Arabic company to create software programs to facilitate Arabic MT and other programs for the natural processing of Arabic for various reasons. Al-Alamiyyah Group did develop a set of Arabic MT software, but it has not been extensively used due to the subscription fees required to access it.

Farhat and Al-Taani (2015) have noted that there has only been "a minimal amount of work done on the Arabic language as a target language" in the prior literature on MT. Additionally, Western-developed methodologies have been utilized in Arabic research, which may not be appropriate given Arabic's unique traits.

Z. M. Almahasees (2018)conducted a study on the capacity of Google Translate and Microsoft Bing Translator to provide accurate English translations of Arabic-language journalistic materials assessed in this paper. To achieve this, the study has incorporated linguistic error analysis. The study's findings demonstrate that both algorithms produce excellent orthography and grammatical results with > 90% accuracy. Moreover, lexical and grammatical collocations yield positive results for both systems, with 79.8% for Google and Microsoft Bing.

Z. M. Almahasees (2017) has chosen the automatic evaluation of the two system outputs using the most widely used automatic evacuation metric, BLEU, to conduct the investigation. The study's corpus consists of 25 Arabic sentences taken from the

DOI: http://doi.org/10.24086/ICLANGEDU2023/paper.931

Petra News Agency of Jordan and translated into English using human references. The study's findings demonstrated that Google Translate outperforms Microsoft Bing when compared to human-referenced translation.

Zakraoui, Saleh, Al-Maadeed, and AlJa'am (2020) reviewed three Machine Translation approaches used by MT systems that offer translation from English into Arabic translation. The study found that the attention-based approach, which was recently developed, does better than the previous approaches. Ali (2020) conducted a survey evaluating three MT systems: Google Translate, Microsoft Bing, and Ginger. He assessed the translation of the UN documents from English into Arabic regarding fidelity and intelligibility errors. He found that Microsoft Bing achieved the best performance, followed by Ginger and Google Translate. Finally, Taleghani and Pazouki (2018) conducted a study to verify MT's capacity to translate English proverbs into Persian. Their study found that Tarjamn software is better than other systems for rendering English proverbs into Persian.

Even though there have been numerous studies on English-to-Arabic machine translation systems, the improvement of any widely used and free multilingual MT system for Arabic did not result from these studies. Furthermore, due to a lack of academic collaboration, most of the recommendations from earlier studies have not been implemented. Moreover, language feedback was not considered in the earlier studies, which approached MT research from a computational standpoint.

With the lack of linguistic analysis for MT, the current work attempts to fill the gaps in the limited English-to-Arabic MTE. The study's corpus, evaluation procedures, data analysis, and methodology will all be covered in detail in the following section.

III. Methodology

A. Corpus

The United Nations Security Council, the League of Arab States, Petra News Agency, and the World Health Organization are just a few international organizations whose websites were used to compile the study's corpus. The corpus size of the study consisted of 400 sentences extracted from the websites of the above organizations. As was mentioned above, our parallel bilingual corpus for generating translational equivalents uses both the original texts of the corpus and their cited translation into either English or Arabic. These texts were chosen primarily for their ability to offer a rich domain in register collocations. Each text has register collocations peculiar to a given field, making them perfect for testing and assessing the MT's potential. In addition, this variety of texts is primarily intended to attempt objectivity in evaluating translation systems that use the most cutting-edge translation methodologies, such as Neural Machine Translation and Hybrid Machine Translation.

About 400 sentences worth of selected samples have been chosen from the four domain documents. These excerpts were used as input data to test the three MT systems' ability to handle translation from English to Arabic. The overall translation quality of each MT system was determined by testing the extracts. This research made it possible for the researcher to decide which MT System, compared to others, generally produced superior translation quality for domain documents. The current study, which analyzes the output of the three systems over one year of testing regarding the quality of each system's output, is synchronic. The above data is analyzed using Fidelity and Intelligibility scale (LDC, 2002; Sin-Wai, 2015), as shown in Figure 1.

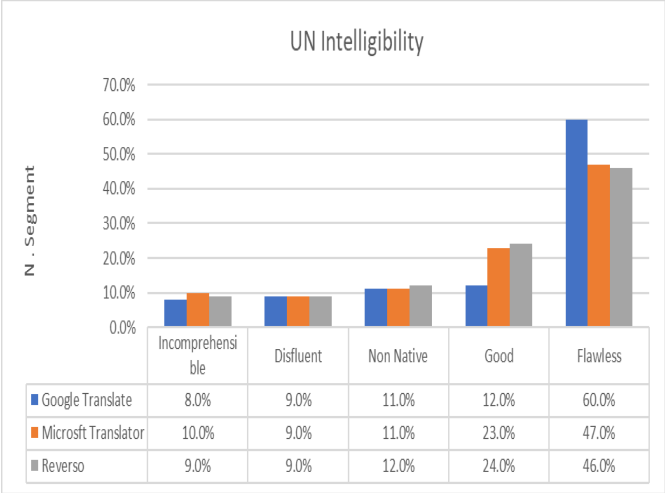
IV. Analysis

This part of the study analyses the chosen corpus using the adopted framework of fidelity and intelligibility scales (LDC, 2002).

A. United Nations

The UN is one of the most significant document generators in the world due to the nature of its operations. For example, the Security Council or General Assembly will always issue UN resolutions and reports, formal writings endorsed by the UN body. The UN Digital Library, which has provided free access to UN publications since 1979, has these documents available for download. Additionally, all UN documents are translated into the organization's six official languages after being written in English.

1. Intelligibility



	Incomprehensible	Disfluent	Non Native	Good	Flawless
Google Translate	8.0%	9.0%	11.0%	12.0%	60.0%
Microsft Translator	10.0%	9.0%	11.0%	23.0%	47.0%
Reverso	9.0%	9.0%	12.0%	24.0%	46.0%

Figure 2. UN Intelligibility

The above figure shows how the three systems have rendered the UN information into English and to what degree the output of the three systems is intelligible. The above results indicate that the three systems have achieved a gradual improvement. It shows that the highest intelligibility score was ***flawless***, where the output is well grammatically formed. Google Translate achieved the highest percentage of ***flawless*** output with 60%, followed by Microsoft Translator with 47%, and then 46% for Reverso. Reverso system rendered the UN systems with ***good*** scale with minor grammatical errors; Reverso outperformed the other systems with 24%, followed by Microsoft Translator with 23%, and lastly, Google Translate with 12%. For the Non-Native scale, Google Translate and Microsoft Translator achieved the minimum number of non-native outputs at 11%

DOI: http://doi.org/10.24086/ICLANGEDU2023/paper.931

for each system, followed by Reverso with 12%. On the other hand, the three systems achieved a similar percentage of disfluent output, which contains a lot of grammatical mistakes, with 9% for each system. Finally, Google Translate obtained the least incomprehensible outputs with 8%, followed by Reverso with 9%, and then Microsoft Translator with 10%.

2. Fidelity

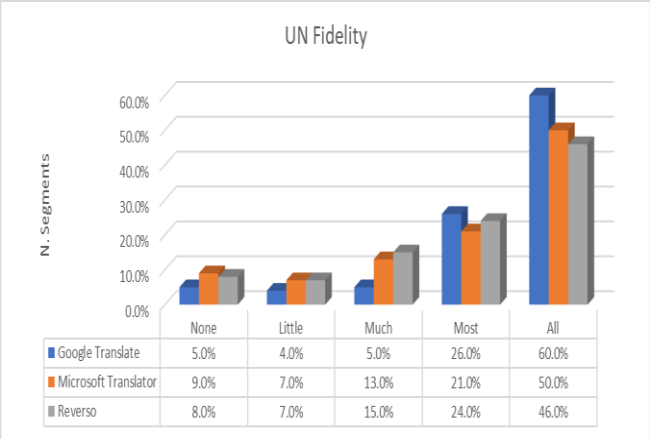


Figure 3. UN Fidelity

The above figure illustrates how the three systems have rendered UN documents regarding Fidelity. Google Translate achieved higher performance than the systems in terms of ***All*** scales. The system produced fluent output without any grammatical flaws with 60%, followed by Microsoft Translator with 50% and then Reverso with 46%. Then Google Translate also outperforms other systems in providing output with minor fidelity errors for 26 sentences, equal to 26%, followed by Reverso with 24 sentences, equivalent to 24%, and then 21 sentences for Microsoft Translator. On the other hand, Google Translator achieved the lowest number of outputs with several grammatical errors at 5%, followed by Microsoft Translator at 13% and then Reverso with 15%. Moreover, Google Translate also surpassed the other systems in providing the lowest number of sentences with several grammatical structures, with one fragment of the output with grammatical errors, followed by Microsoft Translator and Reverso with 7% for each system.

B. Arab League

Arab League is the most prominent regional organization. It promotes cooperation among Arab nations to protect their autonomies and sovereignty and the interests of the Arab countries. Periodically, the League of Arab States publishes its documents and resolutions in Arabic and sometimes in English.

1. Intelligibility

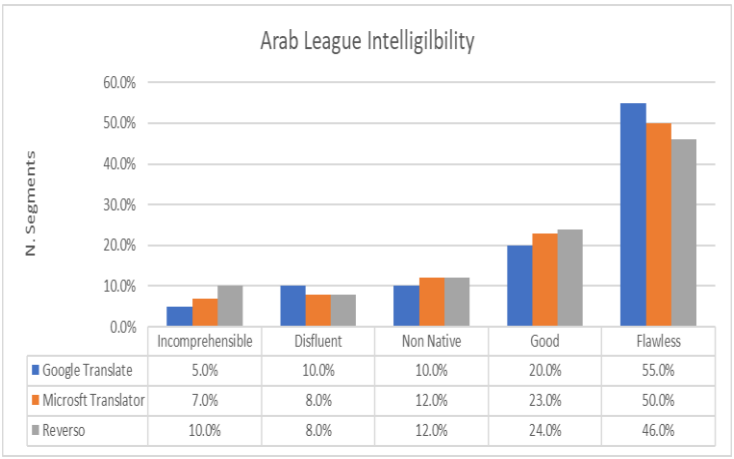


Figure 4. Arab League Intelligibility

The accompanying graph demonstrates how well the three selected systems translate Arabic texts into English. It indicates that the flawless score receives the maximum number of Intelligibility scores, while incomprehensible scores receive the lowest number. Google Translate outperformed the three systems in providing the highest number of English outputs without grammatical mistakes at 55%, followed by Microsoft Translator at 50% and then Reverso with 46%. Then, Reverso achieved the highest number of English outputs with minor mistakes at 24%, followed by Microsoft Translator at 23%, and then Google Translate Lastly with 20%. For the Non-native scale, Google Translate also provided the lowest number of non-native English output with 10%, followed by 12% for Microsoft Translator and Reverso. Microsoft Translator and Reverso produced the lowest number of disfluent sentences for the *disfluent* scale. Like disfluent, Google continues to obtain the lowest number on the incomprehensible scale, where the output is difficult to understand at 5%, followed by Microsoft Translator with 7%.

2. Fidelity

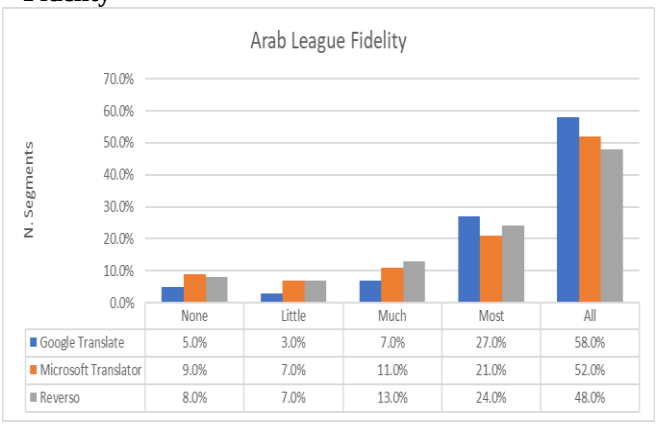


Figure 5. Arab League Fidelity

The fidelity levels attained by the three systems in handling Arabic texts in English are shown in the above chart. It demonstrates that the score for everything, where the ST meaning is fully stated without any addition or deletion in the TT, receives the highest percentage throughout assessing the three systems. Google Translate achieved the highest rate of

DOI: http://doi.org/10.24086/ICLANGEDU2023/paper.931

rendering Arabic output into English with 58%, while Reverso achieved the lowest number of sentences free of any fidelity errors, equal to 48%. For Most scales, Google Translate reached the maximum number of Arabic sentences into English with minor adequacy errors. As a result, Google Translate occupied the first rank for the rest of the fidelity rankings, followed by Microsoft Translator and then Reverso.

C. *Petra News Agency*

Petra News Agency is the official news agency of Jordan. One of the top objectives for accurately reporting what is happening in Jordan and other parts of the world is translating journalistic texts.

1. Petra Intelligibility

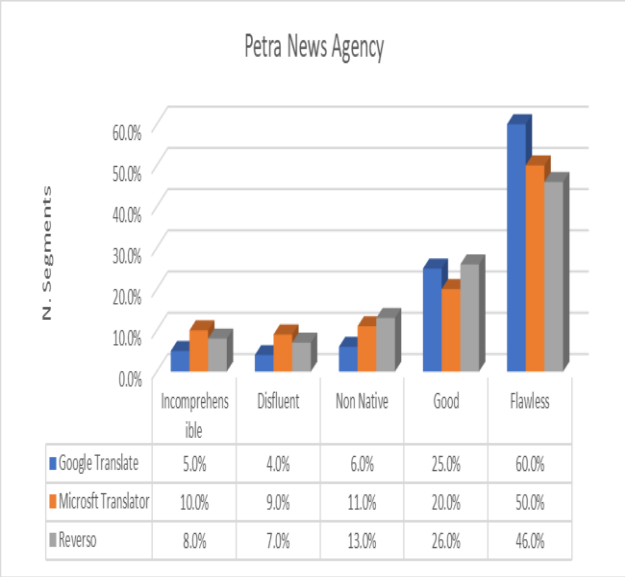


Figure 6. Petra News Agency Intelligibility

The adequacy of the three systems in providing journalistic texts from Arabic to English is illustrated in the above chart. The results showed that Google Translate outperformed the three systems, followed by Microsoft Translator and Reverso interchangeably.

2. Fidelity

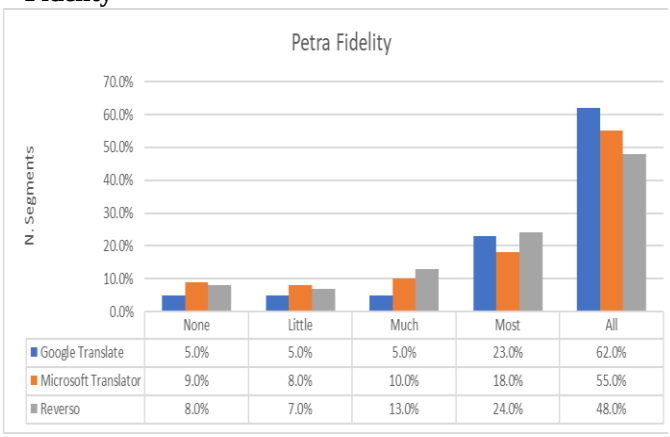


Figure 7. Petra News Agency Fidelity

The chart shows the level of Fidelity attained by the three systems when translating Arabic journalistic texts into English. The ***All*** score receives the maximum fluency rating, while the None score receives the lowest rating due to the output's poor quality and difficulty in comprehension. Google Translate outperforms the other systems in providing the highest number of All scores, which equals 60% of the text, followed by Microsoft Translator, which represents 55%, and then Reverso with 48%. Regarding most scores, Microsoft Translator provides the minimum number of the most score with 18%, which classifies it first, followed by Google Translator with 23% and then Microsoft Translator with 24%. On the other hand, Google Translate occupies the lowest number of errors with 5 % for each scale, and then Microsoft Translator and Reverso interchangeably followed Google Translate.

D. *WHO*

The World Health Organization is a specific UN organization that oversees global public health (WHO). Working internationally to advance health, uphold international security, and aid the weak is required by the WHO's mandate.

1. Intelligibility

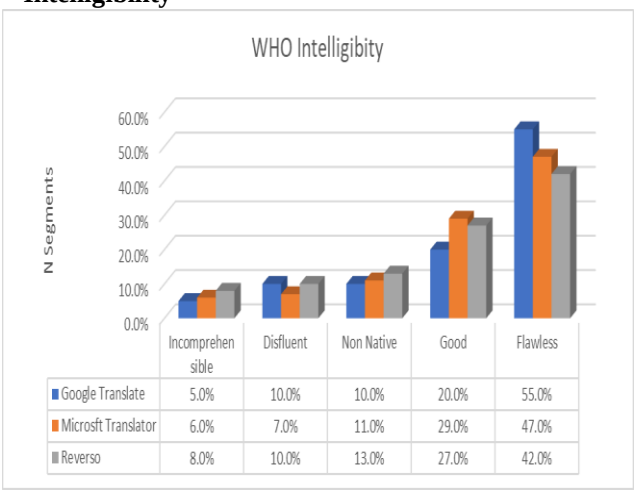


Figure 8. WHO Intelligibility

The above graph illustrates the degree of intelligibility made by the three systems. Google Translate surpasses the other systems in translating the maximum number of flawless scores with 55 segments, equaling 55%, followed by Microsoft Translator with 47 *flawless* segments, bringing a total of 47%. Then Reverso with 42 *flawless* sentences, equaling 42%. Microsoft Translator outperforms the other three systems by providing *good* scale translation with minor errors, equaling 29%, followed by Reverso with 27%, and then Google Translate. On the other hand, Google Translate outperforms the different scales in providing the lowest numbers of errors, followed by Microsoft Translator and Reverso interchangeably.

2. Fiedility

DOI: http://doi.org/10.24086/ICLANGEDU2023/paper.931

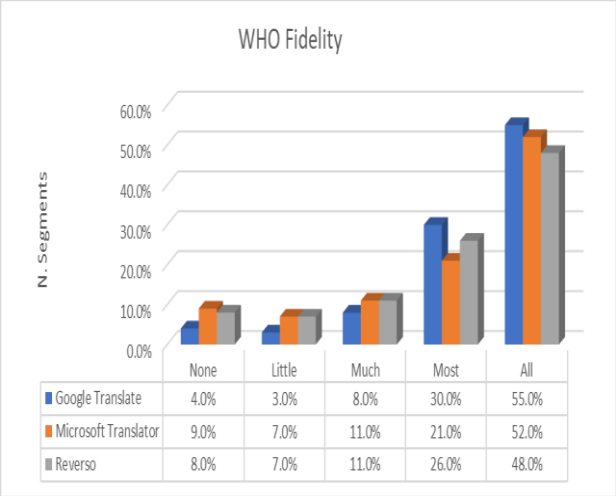


Figure 8. WHO Fidelity

The above chart illustrates the degree of Fidelity attained by the three systems. Google Translate outperformed the other systems by providing All scale sentences, followed by Microsoft Translator and Reverso. For Most scales, Google Translate continues to surpass the other systems in providing translation with minor meaning errors, equalling 30%, followed by Reverso with 26% and then Microsoft Translator with 21. On the other hand, Google Translate also outperformed the other systems, followed by Microsoft Translator and Reverso interchangeably.

Conclusion

Since NLP and MT were primarily created to aid end users in carrying out translation duties across languages, they have acquired the utmost relevance. With many MT services accessible, including translating papers, speech, and images, MT is becoming more widespread in all walks of life. These services are now readily available, which opens the door for several review activities to confirm the acceptability of MT. The current study compares three multilingual free MT systems, Google Translate, Microsoft Translator, and Reverso, each offering an Arabic translation service. The study showed that Google outperformed the other systems in providing translation for both English and Arabic in terms of fidelity and intelligibility scales. Regarding intelligibility and fidelity outcomes throughout the evaluation, Google Translate ranked first in the translation of UN and WHO, Petra News Agency, and Arab, followed by Microsoft Translator in second place and Reverso in third place.

References

Adly, N., & Al Ansary, S. (2009). *Evaluation of Arabic machine translation system based on the Universal Networking Language.* Paper presented at the International conference on application of natural language to information systems.

Ali, M. A. (2020). Quality and machine translation: An evaluation of online machine translation of English into Arabic texts. *10*(5), 524-548.

Almahasees. (2021). *Analysing English-Arabic Machine Translation: Google Translate, Microsoft Translator and Sakhr*: Routledge.

Almahasees, Z. (2021). *Analysing English-Arabic Machine Translation: Google Translate, Microsoft Translator and Sakhr*: Routledge.

Almahasees, Z., Husienat, I., & Husienat, A. (2022). Perceptions of University Students Toward Blended Learning During COVID-19. *22*(18).

Almahasees, Z., & Jaccomard, H. (2020). Facebook translation service (FTS) usage among jordanians during COVID-19 lockdown. *5*(6), 514-519.

Almahasees, Z. M. (2017). Assessing the translation of Google and Microsoft Bing in translating political texts from Arabic into English. *3*(1), 1-4.

Almahasees, Z. M. (2018). Assessment of Google and Microsoft Bing translation of journalistic texts. *4*(3), 231-235.

Altintas, K., & Cicekli, I. (2022). *A machine translation system between a pair of closely related languages.* Paper presented at the Proceedings of the 17th International Symposium on Computer and Information Sciences.

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *26*(2), 137-161.

Dabre, R., Chu, C., & Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *53*(5), 1-38.

Farghaly, A., & Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *8*(4), 1-22.

Farhat, A., & Al-Taani, A. (2015). *A rule-based English to Arabic machine translation approach.* Paper presented at the The International Arab Conference on Information Technology (ACIT'2015).

GoogleTranslate. (2022). Google Translate.

LDC. (2002). Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations.(2002). In.

MicrosoftTranslator. (2022). Translator language support.

Reverso. (2022). Reverso Translation.

Salem, Y., Hensman, A., & Nolan, B. (2008). *Implementing Arabic-to-English machine translation using the Role and Reference Grammar linguistic model.* Paper presented at the Proc. 8th Annu. Int. Conf. Inf. Technol. Telecommun.(ITT).

Sin-Wai, C. (2015). The Routledge encyclopedia of translation technology.

Taleghani, M., & Pazouki, E. (2018). Free online translators: a comparative assessment in terms of idioms and phrasal verbs. *6*(1), 15-19.

Zakraoui, J., Saleh, M., Al-Maadeed, S., & AlJa'am, J. M. (2020). *Evaluation of Arabic to English machine translation systems.* Paper presented at the 2020 11th International Conference on Information and Communication Systems (ICICS).

DOI: http://doi.org/10.24086/ICLANGEDU2023/paper.931