University of Tehran press

JOURNAL OF FOREIGN LANGUAGE RESEARCH

p-ISSN:2588-4123 e-ISSN:2588-7521

https://jflr.ut.ac.ir Email:jflr@ut.ac.ir

On Microsoft Translator's Performance in English-Persian speech-to-text Translation: Recognizing Translation Errors and Identifying their Sources

Pardis Qasemi ✉* 0000-0002-0114-2488

Department of Foreign Languages, Faculty of Literature and Humanities, Shahid Bahonar University of Kerman, Kerman, Iran..

Email: pardis.qasemi1995@yahoo.com

Sima Ferdowsi ** 0000-0003-4556-5112

Department of Foreign Languages, Faculty of Literature and Humanities, Shahid Bahonar University of Kerman, Kerman, Iran.

Email: sima.ferdowsi@uk.ac.ir

Najme Bahrami *** 0000-0002-0114-2487

Department of Foreign Languages, Faculty of Literature and Humanities, Shahid Bahonar University of Kerman, Kerman, Iran.

Email: bahrami.n@uk.ac.ir

ABSTRACT

Over the past decades, the language industry has benefited from computer-aided tools. Although these tools has not affected interpreting to the same extent as translation, but some improvements have been made in interpreting as well. The desire to avoid cognitive saturation has increased interest in computer-assisted tools and speech translation systems among interpreters. However, to better understand the function of these systems, it is necessary to examine their output, identify possible errors, and evaluate output data quality. Despite worldwide interest in detecting the role of technology in interpreting, it seems that this research area has been quite under-researched in Iran. In an attempt to fill this research gap, the present descriptive study aimed to investigate the performance of Microsoft Translator. The researchers intended to identify the output created by this software, detect the errors and their probable sources with the goal of utilizing Microsoft Translator as an assistant tool in interpreting classes based on its performance. To that end, corpora of hearings addressed at United Nations sessions, their speech-to-text translations by Microsoft Translator, and a reference translation were collected and analyzed. To find answer to the first research question, Microsoft Translator errors were detected and categorized based on the component responsible for generating the errors. The MT-based errors were classified based on the taxonomy of Costa et al. (2015). The ASR-based errors were also recognized and categorized. The second question concerned the probable causes of errors. The findings showed that Internet access, time delay, manual function of the microphone, and speaking features could lead to translation errors. The findings of this research can be a starting point for future research in the field of computer-assisted tools in different modes of interpreting. Moreover, evaluating the performance of translation assistance tools between different language pairs can assist the creators and designers of these tools in improving and enhancing these systems.

ARTICLE INFO

Article history:

Received: 08 April 2023

Received in revised form 31 July 2023

Accepted: 04 August 2023

Available online: Autumn 2023

Keywords:

Computer-assisted Interpreting; Conference Interpreting; Error Analysis; Speech Translation Systems; Microsoft Translator

Qasemi, P.; Ferdowsi, S., & Bahrami, N. (2023). On Microsoft Translator's Performance in English-Persian speech-to-text Translation: Recognizing Translation Errors and Identifying the Source of Errors. *Journal of Foreign Language Research*, 13 (3), 439-456. http//doi.org/ 10.22059/jflr.2023.357547.1028

 Publisher: The University of Tehran Press.

DOI: http//doi.org/ 10.22059/jflr.2023.357547.1028

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 13, Number 3, Autumn 2023, Page 439 to 456

✉* Pardis Qasemi , Department of Foreign Languages, Faculty of Literature and Humanities, Shahid Bahonar University of Kerman, Kerman, Iran

** Sima Ferdowsi, Department of Foreign Languages, Faculty of Literature and Humanities, Shahid Bahonar University of Kerman, Kerman, Iran

*** Najme Bahrami, Department of Foreign Languages, Faculty of Literature and Humanities, Shahid Bahonar University of Kerman, Kerman, Iran

1. Introduction

Psychologists, linguistics, and interpreters admit that simultaneous interpreting (SI) is a challenging cognitive task involving a fundamental psycholinguistic process (Al-Khanji et al., 2000). To perform this task, the interpreter must constantly track, save, and retrieve the source language input to provide the target language output orally. According to Mohammadi (1401: 135), "in the process of simultaneous interpreting, the interpreter concurrently engages in a creative, active and linguistically dynamic approach to deciphering and encoding information in the source and target languages". Consequently, SI is a demanding cognitive and linguistic activity, even for expert interpreters, which forces them to look for lexical or syntactic search strategies. Yet, SI difficulty is not all about the linguistic skills and mastery of the source and target languages. It also involves extra-linguistic knowledge and being updated on the topic of a particular task. Interpreters, depending on the field they are working in, are faced with different threats.

Interestingly, Gile (1995) in his study concluded that some of the mistakes found in simultaneous and consecutive interpreting output could not be readily a result of insufficient linguistic knowledge, lack of extra-linguistic skill, or terrible conditions of source speech delivery. To explain the reasons behind such interpreting errors, Gile (1995) introduced two separate models for simultaneous and consecutive interpreting, where the practice of interpreting is analyzed into several components. The components were called 'efforts' to emphasize the burdensome nature of interpreting. The 'Effort Model' assembles the operational components of SI into four efforts namely the listening and analysis effort (L), the production effort (P), the short-term memory effort (M) and a coordination effort (C). According to this model, to ensure an acceptable interpreting, the following equality should be hold:

I. SI = L+ M+ P+ C

If the required effort for SI exceeds the available capacity of the interpreter, saturation occurs, resulting in errors, omissions, or inadequacies in the interpretation. Moreover, any imbalance in the efforts of simultaneous interpretation can lead to incorrect translations. For example, interpreters who devote too much processing capacity to the Production Effort, with a desire to produce an eloquent output, would "end up with insufficient processing capacity for the Listening and Analysis Effort." (Gile, 1995: 175)

The underlying principle of the Effort Models is the fact that, on the one hand, the amount of energy required for interpreting is limited in supply. On the other hand, almost all mental energy is sometimes used, and the interpreter may need even more to perform its duty. Consequently, the imbalance between the available mental energy and the energy required for interpretation negatively affects the quality of the task. In other words, the interpreter is forced to disrupt the balance between efforts. Additionally, some features of the source speech, such as speech density, unusual accent, and syntactic differences between the source and target languages, require the simultaneous interpreter to exert more effort.

Given the particular difficulties of SI, finding solutions to facilitate this process is essential. Technology is one of the tools available to humans in all aspects of life. Translators also benefit from the positive effects of technology in

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 13, Number 3, Autumn 2023, Page 439 to 456

their language activities, and during the past decades, the language industry has benefited from computer-assisted translation tools. Although the impact and use of information technology in the field of SI is not as significant as its impact on written translation, it is still very effective in the performance of simultaneous interpreters. For example, through web data, interpreters can access the necessary information from around the world (Fantinuoli, 2018). In addition, the use of laptops, tablets, and digital devices facilitates the search for various terms during simultaneous interpretation or conference sessions (Tripepi Winteringham, 2010).

In the current market where demand for simultaneous interpreters is increasing day by day, one possible solution to meet the market needs and ensure interpreting quality is the use of technology in interpreting. Therefore, attempts have been made to investigate the performance and efficiency of the Microsoft Translator in SI, with an emphasis on the use of translation assistance tools. This considered the following research questions were posed:

1- What are the main errors generated by Microsoft Translator in SI?

2- What are the probable sources of errors?

2. Literature Review

A review of relevant research reveals that, research on technology and interpreting are divided into two main categories depending on the subject and target they are concerned with; evaluation or assessment and deployment or improvement. In the following section, a few examples of each of these types of research will be presented separately:

2.1. Studies on the Evaluation of Speech Translation Systems

Seligman can be referred to as one of the first researchers who studied speech translation technologies. Seligman (2000) drew up a sketch of six problems related to ST, summarizing his previous studies. The six items examined in this research were interactive disambiguation, system architecture, the interface between speech recognition and analysis, natural pauses for segmenting utterances, dialogue acts, and the tracking of lexical co-occurrences. Seligman enhanced his work by discussing nine issues in the field; data structures, example-based MT, and resolution of translation mismatches were the three issues added. The purpose of these papers was to clarify the various dimensions of ST.

Nakamura et al. (2006) noted the ST barriers between Western and non-Western languages, such as linguistic divergence, word separation, and word order transformation. Their evaluation showed that the strategy obtained was appropriate for high-quality system construction. The study demonstrated that the system needs improvement to translate longer and more natural sentences. However, the adoption of neural systems in machine translation contributed to the growth of ST systems.

Hamon et al. (2009) claimed that the main disadvantage of automatic speech translation systems compared to human interpreters was the translation quality. However, the available state-of-the-art systems could provide understandable translation. In their study at the Karlsruhe Institute of Technology (KIT), they conducted an end-to-end evaluation of a simultaneous speech translation system made up of state-of-the-art components to translate to demonstrate the advantage of such systems. The findings showed that one advantage of ST systems compared to human interpreters was a short memory, which made the system independent of compensatory strategies. Another feature of the ST system which made it favorable was being cost-efficient. Once the system was designed and adapted to the target domain, it could be reused several times. Working with human interpreters was costly as it required hiring two interpreters each time, sound proofing the booth, and providing audio equipment.

In the recent past, researchers focused their attention on presenting prototype systems according to particular societies' needs. For instance, the Karlsruhe Institute of Technology (KIT) attracts students from all over the world. The fact that classes were held in German was one of the problems international students at KIT faced. To solve this problem, Müller et al. (2016a) proposed a Lecture Translation System (LTS) for KIT. Later on, Müller et al. (2016b) conducted another study to verify the quality of LTS. To this end, they distributed a questionnaire to students who had spent two semesters working with LTS. The findings revealed that LTS was beneficial to students, and the researchers were also able to identify the system's key weaknesses to be improved.

The advantages and disadvantages of using technological tools and systems have always been debated. One of the drawbacks of interpreter-assisted devices and systems is that they are handled manually. The findings of empirical studies on the use of CAI tools seem to prove the idea that interpreters may have the time and the cognitive capability to manually look up specialized terms while they are in the booth (Prandi, 2015; Biagini, 2016). However, an automated querying system could reduce the required cognitive effort. Therefore, Fantinuoli (2017) proposed making the process automated by combining ASR and CAI tools to reduce the burden of additional mental effort imposed on the interpreting process due to manual handling of CAI tools.

In a recent study, Almahasees (2018) adopted an error analysis method to evaluate the translation capacity of two systems, Google Translate and Microsoft Bing. The findings showed 90% accuracy in orthography and grammar due to the adoption of neural machine translation. Both systems showed more than 79% accuracy in operation, considering lexical and grammatical collocations. Although the study did not examine the function of ASR and transcription, it provided useful information on Machine Translation (MT), one of the components of ST systems.

2.2. Studies on the Improving the Performance of Speech Translation Systems

In the previous section, we reviewed evaluative researches. This section introduces the studies that have applied speech translation systems in different domains and different perspectives to investigate ST's features and drawbacks. According to Valipour (2021: 553), "the success of machine translation in the future depends on the effort to recognize and extract semantic relationships and syntactic structures, categorize them, and formulate them for application in this field."

The error-prone nature of current ST technologies is among their shortcomings. To improve the ST System's application's accuracy, Frederking et al. (2000) offered interactive error correction by users. Thus, the Multi-Engine Machine Translation (MEMT) architecture was designed so that errors could be corrected interactively throughout the system.

Waibel et al. (2003) designed a prototype speech to speech (STS) system in the domain of medical interviews. The system concurrently translates the conversation between an English language doctor and Egyptian Arabic patients. Although this prototype was limited to only hundreds of sentences in English and Arabic, the result demonstrated the feasibility of such a system. The system could provide translation with 80% accuracy and time lag of 2-3 seconds. However, one of the weaknesses of this system was the reduction in the quality of sound input recognition in noisy environments.

The progression of STSs has become evident through numerous studies demonstrating these systems' ability to translate from and to different languages. However, as languages and domains are continually expanding, concerns about the power of STSs to keep up with ever-changing domains have been raised. Schultz et al. (2006) developed an STS with a specific architecture to solve system maintenance and data insufficiency problems. Initially, this system was tested with the English/Thia Doctor-patient scenario. Furthermore, the application of this system was re-evaluated in real-life scenarios of medical conversations for the same language pairs. The results showed that some features of this system, such as the provision of audible feedback and output, microphone accessibility, and the use of push-to-talk, were invaluable to users.

Along with other fundamental goals for developing speech translation systems, such as reducing costs or increasing the translation quality, it is to help people with particular disabilities. Therefore, Singh and Singh (2014) presented a text-to-speech (TTS) translation system which could translate English to Punjabi conversation. Not for the mere sake of translation but to assist visually impaired people.

3. Theoretical Framework

Interpreting Studies, according to Roy and Metzger (2014: 158), can be studied "from a variety of disciplines-sociology, anthropology, psychology, linguistics and/or a mix of these disciplines". Accordingly, the cross-disciplinary nature of IS leads to the application of theories, methodologies, and frameworks from more than one discipline. The present study followed an interdisciplinary approach to research between Interpreting Studies, Psychology, and Information Technology.

The research investigated the possibility of applying ST systems to aid simultaneous interpreting. However, given the cognitive demand nature of SI, it is suspected that the application of computational devices may add to the strain on the back of an interpreter already undertaking a challenging job, leading to cognitive saturation. However, the application of ST systems during SI is justifiable considering Seeber's (2011) Cognitive Load Model.

Seeber's (2011) Cognitive Load Model (CLM) of simultaneous interpreting was developed based on Wickens' (1984) Multiple Resource Model (MRM), which explains the cognitive processes in multitasking activities. The model applies to SI since "simultaneous interpreting is an instantiation of multitasking that requires the interpreter to engage in a language comprehension task and a language production task at the same time" (Seeber, 2011: 187). MRM argues that "the combination of two (or more) tasks require more processing capacity than either (or any) of the tasks performed individually" (Seeber, 2011: 187). Moreover, the model implies that tasks with the same level of processing dimension interfere with each other more strongly than tasks relying on different structures. For instance, it is easier and more efficient to perform visual and auditory tasks concurrently than perform two visual tasks. The underlying processes are the same in the latter but are not shared in the former (Wickens, 2002).

Seeber (2007) adopted Wickens' (2002) Model to SI and referred to it as the Cognitive Resource Footprint (CRF). As mentioned before, SI includes two main tasks; the first task is listening and comprehension and the second one is production and monitoring. As presented in CRF, listening and comprehension require auditory-verbal and cognitive-verbal resources at the perceptual-cognitive stage. Production and monitoring, however, need the resources mentioned above for the perceptual-cognitive stage as well as vocal-verbal resources at the response stage when the message is delivered verbally while at the same time the translation is being checked.

To illustrate the interference between two tasks conducted concurrently, CRF is incorporated into a matrix of conflict. If two sub-tasks share resources with the same structure, their degree of interference is more significant than two sub-tasks sharing different resources, which explains multiple tasks at the same time. Based on this, it is possible to explain the performance of multiple concurrent activities.

The Cognitive Load Model can justify the employment of an ST by the interpreter during the SI task since this model is based on Wickens's (1984) multiple resource theory and assumes that resources can be reallocated among different comprehensive and verbal tasks. In other words, unlike the single resource theory (Kahneman, 1973), which assigns one undifferentiated pool of source to interpreting processes, the multiple resource model, believes in numerous resources that can be shifted between and during different tasks in interpreting. Moreover, the literature supports that CAL tools can assist interpreters during their jobs without causing cognitive saturation. For example, Prandi (2018) showed that not only using the CAI tools did not cause saturation, but they also helped prevent it by reducing local cognitive load.

Considering this framework, the application of CAI tools or ST systems is governed by visual-verbal resources and does not interfere with the underlying structure of other interpreting tasks. Though it may not be easy for beginner interpreters to use these devices at first attempts when performing interpreting, they will get used to it through training and practice. Consequently, they will provide high-quality interpreting

4. Research Method

The present descriptive study, following Williams and Chesterman's (2002: 49) comparative model, adopts a “static and product-oriented approach” to examine the output of Microsoft Translator. To collect the necessary data, speeches delivered at the United Nations were used to create the required corpus. According to Tognini-Bonelli (2001: 55), a corpus is "a computerized collection of authentic texts amenable to automatic or semiautomatic processing or analysis." Corpus-based research uses technology to analyze large collections of electronic texts selected according to the explicit criteria. Straniero Sergio and Falbo (2012) believed that the primary goal of corpus linguistics is to describe numerous dimensions of language, relying on information technology capacities, especially in terms of the arrangement of a significant amount of data.

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 13, Number 3, Autumn 2023, Page 439 to 456

4.1. Corpus Collection Method

As with Corpus-based Interpreting Studies (CIS), the first step in data collection is corpus creation. As mentioned before, corpus-based studies require the compilation of data happened in natural setting. According to Russo et al. (2018), several international organizations such as the European Parliament (EP), the European Commission (EC), and the United Nations (U.N.) can be used as sources of data collection. This is due to the nature of corpus-based studies, which rely on authentic data occurring in a natural environment. For the purpose of the present study, the following steps were taken to gather the intended corpora.

Step one: the researchers downloaded three political videos in English from the U.N. conference website https://www.youtube.com/.

Step two: the transcription of each speech was downloaded from https://www.whitehouse.gov/ and carefully checked with the videos. These texts were provided to the translator in the next stage to prepare a reference translation. The transcribed texts of the three English video files contained 5877 words (Table 1).

NO.	Source Speech Video	Video Duration	Addresser	Language	Transcribed words No.
1	75th U.N. session, 2020	7 minutes	Donald Trump	English	968
2	74th U.N. session, 2019	36-37 minutes	Donald Trump	English	3883
3	71st U.N. session, 2016	9 minutes	Barak Obama	English	1026

Table 1: Description of Source Data

Step three: the Microsoft Translator application was downloaded from https://www.microsoft.com/enus/translator/apps/ and installed on the iOS device and run. It is a free translation app for more than 70 languages that can translate texts, conversations, voices, camera shots, and screenshots. The Microsoft translator is a web-based app; therefore, the device needs internet access. The Microsoft Translator saves all the transcribed and translated segments within the application history. Therefore, it is accessible for further usage and analysis.

Step four: to identify the translation errors of Microsoft Translator, it was necessary to compare these translations with a reference translation. To this end, the texts produced in the second stage were provided to a translator to generate the reference translation by translating them. The translator was a graduate of English translation (with a master's degree) and had eight years of experience as a freelance translator of political texts. The translator had the opportunity to use a dictionary or other CAT tools without any time constraints to complete the translation work.

5. Data Analysis

In order to analyze Microsoft Translator errors, there is a need to identify and categorize errors based on an objective framework. Additionally, the quality of reference translations should also be examined. Below, we will provide detailed and separate explanations for each of these stages.

5.1. Evaluating the Performance of Microsoft Translator

There are two main components in an ST architecture, an ASR and an MT. In some studies, researchers investigate the ST systems as a whole system without differentiating between ASR and MT. In other words, the performance of Microsoft Translator is the result of the performance of these two components.

In some studies, researchers examine speech-to-text translation systems as a complete system without considering the differences between these two components. When ST is evaluated as an entire system, it is not possible to determine which component is responsible for a specific error. To avoid this problem, both ASR and MT output is analyzed and investigated separately in the present study. After the analysis of ASR and MT output separately, the translation provided by Microsoft Translator was compared and contrasted with the reference translation, and the errors were classified according to the taxonomy of MT errors proposed by Costa et al. (2015). This error categorization classifies errors into five categories: orthography, lexis, grammar, semantics, and discourse, each of which includes subcategories.

5.2. Evaluation of Reference Translation

The researchers needed a scoring method to evaluate the quality of the reference translation. For the purpose of the present study, we used the 'Translation Quality Assessment Rubric' introduced by Samir and Tabatabaee-Yazdi (2020). The rubric is a 23-item assessment scale that includes CAT skills in its issue types. Furthermore, the consideration of technology in translation assessment is the strong point of the selected model because it underlines the changes brought into the translation industry due to technological advances and market demands. Therefore, Samir and Tabatabaee-Yazdi's model is in line with the competence framework offered by the EMT Competence Framework (2017), which classified translators' competence into five categories: language and culture, translation, technology, personal and interpersonal, and service provision.

Samir and Tabatabaee-Yazdi (2020) used the Rash measurement model to validate the rubric, which proved that it "had an acceptable person separation reliability of .67 and item separation reliability of .96" (Samir & Tabatabaee-Yazdi, 2020: 117). In this rubric, any translation is assessed on a four-point Likert scale ranging from 'superior= score 4', 'advanced= score 3', 'fair= score 2', to 'poor= score 1'. Therefore, translations with the lowest level of quality got a minimum score of 1. However, score 4 belongs to translations with the highest level of quality.

To make a judgment on translations offered by Microsoft Translator, there needs to have a flawless reference translation. To this end, two raters were asked to score the reference translation to ensure its quality. Both evaluators are university professors with over 15 years of experience in teaching and translating political texts. An interrater reliability analysis using the Kappa statistic was performed to determine consistency among raters. This coefficient is a numerical value between +1 and -1, and the closer it is to +1, the more it indicates the agreement between the different evaluators. As a rule of thumb, values of Kappa from 0.40 to 0.59 are considered moderate, 0.60 to 0.79 substantial, and 0.80 outstanding (Landis & Koch, 1977). The interrater reliability for the translation of 71st U.N. conference delivered by Barak Obama was found to be kappa = .42, p=.03, which is a moderate value; with nearly 45% agreement. The findings showed that for the translation of the 74th U.N. conference delivered by Donald Trump, there was a moderate value for the raters with nearly 55% agreement (kappa = .54, p=.00). Finally, the interrater reliability for the translation of 75th U.N. conference delivered by Donald Trump was found to be kappa= 59, p=.00 (moderate; nearly 60% agreement).

6. Results

6.1. Answer to Research Question No. 1

The first question of the research was about the errors of Microsoft Translator. To answer this question, the reference corpora were compared and aligned with the output corpora of Microsoft Translator, and after detecting the errors, they were classified according to the categorization proposed by Costa et al. (2015). This model provides a comprehensive language classification for written translation.

As stated before, the present study aimed to evaluate the errors produced by both components of Microsoft Translator, i.e., ASR errors and MT errors. Therefore, the output produced by each component was analyzed separately. Accordingly, the data are categorized around MT-based and ASR-based error typologies.

6.1.1 MT-based Errors

Examination of the data revealed a total of 49 MT-based errors. Table 2 presents them separated by type.

Concerning orthography errors, MT's output included 2 spelling errors. Since the Persian language is not sensitive to capital letters, no instance of capitalization error was found. Furthermore, MT performance did not lead to punctuation errors. Orthography errors counted 4.08% of total errors.

Regarding lexis errors, omission, addition, and untranslated word errors were present in MT's translation which counted about 26.53% of total errors. Among different types of lexis errors, untranslated words represent the minority of error.

The most frequently occurring error type was grammar errors which account for 44.89%. Among the subcategories of this error, word order had the least amount of errors, while agreement error was the most common type of error.

24.48% of total errors were counted as semantic errors among which collections had the least occurrence in data. It should be mentioned that the present study focused on the analysis of ideas at the sentence level. Therefore, the alignment was done at the sentence level too. As a result, the fifth type of error, referred to as discourse error in the categorization proposed by Costa et al. (2015), was not examined in the analysis of Microsoft Translator's data.

Error Types		Number of Errors per Type	Total Number of Errors
Orthography Errors	Spelling errors	2	2
	Capitalization errors	0	
	Punctuation errors	0	
Lexis Errors	Omission errors	6	13
	Addition errors	4	
	Untranslated word errors	3	
Grammar Errors	Word class errors	6	22
	Verbal level errors	3	
	Agreement errors	10	
	Contraction errors	0	
	Misordering errors	2	
Semantic Errors	Confusion of sense errors	4	12
	Wrong choice errors	3	
	Collocation errors	1	
	Idiom errors	4	
Total Number of MT Errors			**49**

Table 2. Microsoft Translator Mt-based Errors Type

6.1.2. ASR-based Error

As mentioned before, MT could not be blamed for all types of errors created by the ST system, but automatic speech recognition malfunction is also responsible for error generation. Therefore, evaluating the performance of ASR and discussing ASR-based errors and their consequences on the final performance of Microsoft Translator is essential.

The data analysis led to the identification of 128 errors due to the incorrect performance of the automatic speech recognition system. The results showed that incorrect ASR performance can lead to the occurrence of multiple errors in translation. In other words, the inappropriate functioning of ASR causes MT to produce one of the following errors:

- **Mispunctuation: Referring to the corpus analyzed in this research, the case of punctuation error is a frequent one. In other words, ASR's inability to detect the end of a sentence leads to incorrect use of punctuation marks. For example, instead of ending a sentence with a period, it connects it to the next sentence with a comma. Additionally, in some**

cases, ASR mistakes a declarative sentence for an interrogative one and marks it accordingly.

Example:

Original speech: **…** the China virus __which has claimed countless lives in 188 **countries. In** the United States, …

ASR Recognized speech: *The virus, which has claimed countless lives in 188 **countries in the** United States, ….

MT translation: *این ویروس که جان بی شماری را **در ۱۸۸ کشور آمریکا** گرفته است

As shown in the example, ASR has not recognized that the sentence is finished and merged the two sentences. This incorrect speech recognition has affected the performance of MT and ultimately resulted in an inaccurate translation output.

- Irrelevant addition: In some cases, ASR adds a word or phrase that was not used in the original speech, which results in MT's translation being changed or distorted. This issue usually occurs in pauses between two successive sentences

Example:

Original speech: …….. (no speech was uttered)

ASR Recognized speech: *A nice

MT translation: *یک خوب

As we can see in this example, no sentence was produced in the original language. However, the automatic speech recognition system, for an unknown reason, hears the phrase 'A nice' and registers it, which is then translated by Microsoft Translator.

- Failure to recognize the original input and replacing it with irrelevant words: Sometimes, ASR replaces a word or phrase with another word/phrase, which leads to a complete change in meaning. It seems that one of the reasons for this error is the pronunciation of the original speech.

Example:

Original speech: We reached a landmark breakthrough with **two peace deals** in the Middle East, …

ASR Recognized speech: * we reached a landmark breakthrough with **two piece tails** in the Middle East.

MT translation: *ما با **دو دم تکه** در خاورمیانه به یک پیشرفت برجسته رسیدیم

In general, the findings of the present study regarding the role of speech recognition systems in producing translation errors have also been addressed in other studies. Ruize and Federico (2014), for instance, assessed the influence of Speech Recognition (SR) errors on the quality of MT between English and French. They proposed a statistical framework for their analysis and considered the discrepancy of Automatic Speech Recognition (ASR) systems and the difficulty of utterances in particular settings. The results indicated that SR errors harmed the quality of MT, but it was claimed that different types of errors did not affect the quality of MT to the same degree. Their research results are consistent with the findings of the present study.

6.2. Answer to Research Question No. 2

The second question aimed to identify potential sources of error. The findings showed that the following factors can have a significant impact on the performance of Microsoft Translator and error production:

1. Internet access: Microsoft Translator is an online application. Therefore, it is crucial to access the internet while using it. However, this dependency on the Internet can pose some limitations. Logically, poor internet connection

can slow down the ST process and decrease its effectiveness.

2. Time-lag: Microsoft Translator functions with a short time lag. A few seconds after receiving the audible input, ST transcribes the speech segments into English and, almost concurrently, translates them into Persian, displays it on the device screen. Since this application provides output with a delay, it lags behind in receiving speech input and may miss some parts.

Identifying time-lag as one of the major issues in speech-to-speech translation systems has also been reported in other studies. Fujita et al. (2013), for instance, proposed a method to reduce the delay by dividing the spoken input into smaller units based on pause boundaries. While this approach has been shown to be beneficial for languages with a similar structure, it is less suitable for languages requiring significant word reordering. Moreover, Sridhar et al. (2013), also consider latency in producing output as one of the reasons for errors in translation. They presented segmentation strategies for real-time speech translation of TED talks to balance accuracy and latency. They investigated some methods to improve automatic speech recognition and machine translation quality. Constrained model and vocal length normalization were employed to enhance ASR. MT was also enhanced by adopting monotonic and partial translation retention techniques. The effect of segmentation was observed by inserting different types of text segments. The result proved that the ST system can benefit from a good segmentation.

3. Microphone manual function: Microsoft Translator starts translating when the user manually clicks the microphone button. On the contrary, it will turn off automatically after translating each segment, the length of which cannot be estimated. Therefore, it requires the user's constant attention to turn the microphone on again. The combination of time-lag and the need for manual microphone activation can lead to data omission, which is one of the main problems with this system if it is to be used as an assistance tool to ease the SI process effectively.

4. Speaking features: Another factor that can be considered a probable cause of Microsoft Translator's flaws is the features of spoken data such as accent and long pause. It should be noted that the automatic speech recognition system in Microsoft Translator is of the acoustic type which learns through analyzing hundreds of hours of speech data inputs. Usually, the acoustic model is trained on a limited amount of spoken data which cannot allow for covering all the spoken input variabilities. According to Errattahi et al. (2018), not all speech variables, such as individual differences, accent, vocal differences, and pronunciation, can be taught to the system. Although Microsoft claims that the ASR system is trained based on thousands of hours of input data, there is always the possibility of receiving new input for which the system has not been trained.

7. Discussion and Conclusion

According to Prandi (2020), the new generation of interpreters has a growing interest in computer-assisted interpreting. Several researchers have focused on the usability of CAI tools and their impact on interpreting quality (Biagini, 2016). The results provided a good ground for the inclusion of CAI tools in the curriculum of trainee interpreters. Furthermore, In addition, research conducted in this field can assist software developers and designers in

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 13, Number 3, Autumn 2023, Page 439 to 456

achieving an ideal structure with appropriate performance (Prandi, 2020).

Despite the worldwide interest in detecting the role of technology in interpreting, to our knowledge, the area of CAI tools has been somewhat under-researched in the context of Iran. Therefore, the aim of this study was to fill this gap in academic interpreting research in Iran. To this aim, the study investigated the translation errors produced by Microsoft Translator along the probable sources of errors to determine whether, based on its performance, Microsoft Translator can be used as a CAI tool in SI.

The first question aimed to identify the errors produced by Microsoft Translator. For this purpose, Microsoft's translations were compared with a reference translation, and the identified errors were categorized based on two responsible components. MT errors were classified according to the categorization proposed by Costa et al. (2015). Errors related to ASR were also identified.

The results of the study showed that errors resulting from the performance of MT can be divided into four groups: orthographic, lexical, grammatical, and semantic errors, each with its own subcategories. According to the data from this study, the lowest amount of errors was in the orthographic error group, and the highest number of errors was in the grammatical error group. This finding is not surprising and appears to be due to the structural differences between the English and Persian languages.

Another part of the translation errors resulted from the improper functioning of the automatic speech recognition system. The findings showed that the inappropriate, incorrect, and meaningless translations provided by Microsoft Translator were due to ASR deficiencies. In fact, incorrect recognition of the source input results in the use of incorrect spelling, increased word misplacements, and irrelevant substitutions. Ultimately, these issues lead to incorrect translations by MT.

In response to the second question, data analysis showed that internet access, latency, microphone performance, and speech characteristics can lead to errors in translation. However, it seems that identifying the exact source of the error is not straightforward since these factors are interrelated and can concurrently affect each other and the system's performance. It is possible that an error is due to multiple factors rather than a single factor. For example, ASR may fail to recognize a sentence due to weak internet connectivity and changes in speech language.

According to Hale and Napier (2013), research is a way of acquiring knowledge. However, finding answers to the research question(s) is not the end of knowledge. Instead, answering one question can lead to other questions. The present study is not an exception; while answering the research questions posed for this study, the researchers encountered other questions that could be the starting point for further research:

- Investigating the skills interpreters need to use computer-assisted interpreting tools would be a valuable study.
- Including CAI tools in the process of SI and observing the performance of interpreters will yield interesting results.
- Investigating the use of CAI tools in SI and its impact on the speech disfluency of translators would be another useful research topic.

- As computer-assisted interpreting is a relatively new topic in Iranian universities, there is a need for studies to understand and compare the attitudes of trainee and professional interpreters towards using such systems in translation.
- It is possible to investigate the level of familiarity of Iranian translation students with such systems.
- Another recommendation would be to train interpreters in the use of these systems. Following this, a comparative study could be carried out to determine the difference in quality between simultaneous interpreting using CAI and SI carried out without CAI.

Despite its fascinating findings, the study suffers from certain limitations. The main restriction imposed on this study was that most ST systems are not commercially accessible, and the number of available systems is relatively little. Moreover, not all the free ST systems obtain a good quality to meet the expectation of this study. For the conduction of this research, after exploring the commercially available systems, it was decided to choose Microsoft Translator as the ST system to be studied. The reason behind choosing this system over other systems such as Google Translate was that Microsoft benefits from rich built-up components which provides the users with a wider range of services.

Some restrictions were posed by the Microsoft Translator application itself. First, the Microsoft Speech Translation service is provided for a limited number of language pairs. Regarding the case of this study, ST is available for English to Persian, yet it's not possible to translate speech from Persian to English. As a result, the current study is unidirectional research focusing on translating political hearings from English to Persian.

Another limitation in terms of the Microsoft Translator was the limited capacity of the application. According to experienced gained at the pre-testing phase, it was realized that after about 6 minutes of translation, the application starts to clear the older archived segments. While this problem would not allow saving all ST transcribed and translated segments, it was crucial to have access to all the ST output in order for corpus creation and analysis of data. To solve this problem, the conference videos were broken into parts. Then, instead of playing the video at once, a 5-minute part was played, the saved ST suggestions were copied from the history, and then it was time for playing the next 5-minute part.

Lack of ASR-based error typology can be mentioned as another limitation to the analysis of this study. According to the literature, ST systems are either evaluated as a whole not considering ASR as a separate component, or the ASR evaluation metrics do not investigate the errors flawlessly and in detail. Contrary to the lack of any proposed model for ASR-based errors, since this thesis aimed to analyze the errors generated by ST in terms of its both components (i.e. MT & ASR), attempts have been made to analyze and categorize the ASR errors, apart from MT errors, as detailed as possible.

In conclusion, based on the findings of this study, it can be argued that if access to quality internet is possible, Microsoft Translator could be used as a CAI tool in interpreting classes. However, considering the data capacity limitations mentioned earlier, it seems that this system is more suitable for consecutive interpreting (especially short consecutive

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 13, Number 3, Autumn 2023, Page 439 to 456

interpreting). Certainly, to test this claim and ensure the effectiveness of Microsoft Translator in interpreting classes, this tool needs to be investigated in practice. This requires introducing and teaching the use of this Microsoft Translator to students in interpreting classes.

References

Al-Khanji, R., El-Shiyab, S., & Hussein, R. (2000). On the use of compensatory strategies in simultaneous interpretation. *Meta*, *45*(3), 548-557. Retrieved from https://doi.org/10.7202/001873ar. Accesed on.

Almahasees, Z. M. (2018). Assessment of Google and Microsoft Bing translation of journalistic texts. *International Journal of Languages, Literature and Linguistics*, *4*(3), 231-235. Retrieved from https://doi.org/10.18178/IJLLL.2018.4.3.178. Accesed on.

Biagini, G. (2016). Printed glossary and electronic glossary in simultaneous interpretation: A comparative study [Unpublished doctoral dissertation]. Universita degli studi di Trieste. Retrieved from http://dx.doi.org/10.1075/intp.15.1.04jia. Accesed on.

Costa, A., Ling, W., Luís, T., Correia, R., Coheur, L., (2015). A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation*. 29. 127-161. Retrieved from https://doi.org/10.1007/s10590-015-9169-0. Accesed on.

Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, *128*, 32-37. Retrieved from https://doi.org/10.1016/j.procs.2018.03.005. Accesed on.

European Master's in Translation (2017). Competence Framework. Directorate General for Translation of the European Commission. Retrieved from https://commission.europa.eu/system/files/2018-02/emt_competence_fwk_2017_en_web.pdf. Accesed on.

Fantinuoli, C. (2017). Computer-assisted preparation in conference interpreting. *Translation & Interpreting*, *9*(2), 24-37. Retrieved from https://doi.org/10.12807/ti.109202.2017.a02. Accesed on.

Fantinuoli, C. (2018). Interpreting and technology: The upcoming technological turn. In C. Fantinuoli (Ed.), *Interpreting and technology* (pp. 1–12). Language Science Press. Retrieved from https://doi.org/10.5281/zenodo.1493289. Accesed on.

Frederking, R., Rudnicky, A., Hogan, C., & Lenzo, K. (2000). Interactive speech translation in the diplomat project. *Machine Translation*, *15*(1-2), 27-42. Retrieved from https://doi.org/10.1023/A:1011172330853. Accesed on.

Fujita, T., Neubig, G., Sakti, S., Toda, T., & Nakamura, S. (2013). Simple, lexicalized choice of translation timing for simultaneous speech translation. In INTERSPEECH (pp. 3487- 3491). Retrieved from https://doi.org/10.21437/INTERSPEECH.2013-615. Accesed on.

Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. John Benjamins.

Hale, S., & Napier, J. (2013). *Research methods in interpreting: A practical resource*. A&C Black.

Hamon, O., Fügen, C., Mostefa, D., Arranz, V., Kolss, M., Waibel, A., & Choukri, K. (2009, March). End-to-end evaluation in simultaneous translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL* (EACL 2009) (pp. 345-353). Retrieved from https://doi.org/10.5555/1609067.1609105. Accesed on.

Kahneman, D. (1973). *Attention and effort*. NJ: Prentice-Hall.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174. Retrieved from https://doi.org/10.2307/2529310. Accesed on.

Leeson, L. (2005). Making the effort in simultaneous interpreting: Some considerations for signed language interpreters. In T. Janzen (Ed.), *Topics in signed language interpreting: Theory and practice* (pp. 51-65). John Benjamins. Retrieved from https://doi.org/10.1075/btl.63.07lee. Accesed on.

Mackintosh, J. (1983). Relay interpretation: An exploratory study [Unpublished master thesis]. University of London.

McCarthy, M., & Carter, R. (2001). Size isn't everything: Spoken English, corpus, and the classroom. *Tesol Quarterly*, *35(2*), 337-340. Retrieved from https://doi.org/10.2307/3587654. Accesed on.

Müller, M., Fünfer, S., Stüker, S., & Waibel, A. (2016, May). Evaluation of the KIT Lecture Translation System. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16) (pp. 1856-1861).

Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G. I., Kawai, H., Jitsuhiro, T., ... & Yamamoto, S. (2006). The ATR multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(2), 365-376. Retrieved from https://doi.org/10.1109/TSA.2005.860774. Accesed on.

Pio, S. (2003). The relation between ST delivery rate and quality in simultaneous interpretation. *The Interpreters' Newsletter*, *12*, 69-100.

Prandi, B. (2015). The use of CAI tools in interpreters' training: A pilot study. In Proceedings of the Translating and the Computer 37 Conference (pp. 48-57).

Prandi, B. (2018). An exploratory study on CAI tools in simultaneous interpreting: Theoretical framework and stimulus validation. In C. Fantinuoli (Ed.), *Interpreting and technology* (pp. 29-59). Language Science Press.

Prandi, B. (2020). The use of CAI tools in interpreter training: Where are we now and where do we go from here? *inTRAlinea Special Issue: Technology in Interpreter Education and Practice.* Retrieved from http://www.intralinea.org/specials/article/2512. Accesed on.

Reppen, R. (2010). *Using corpora in the language classroom.* Cambridge University Press.

Roy, C., & Metzger, M. (2014). Researching signed language interpreting research through a sociolinguistic lens. *The International Journal of Translation and Interpreting Research*, 6(1), 158-176. https://doi.org/ti.106201.2014.a09

Ruiz, N., & Federico, M. (2014). Assessing the impact of speech recognition errors on machine translation quality. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track* (pp. 261-274).

Russo, M., Bendazzoli, C., & Defrancq, B. (Eds.). (2018). *Making way in corpus-based interpreting studies.* Springer.

Samir, A., & Tabatabaee-Yazdi, M. (2020). Translation quality assessment rubric: A Rasch model-based validation. *International Journal of Language Testing, 10(2),* 101-128. Retrieved from https://www.ijlt.ir/article_118019.html. Accesed on .

Schultz, T., Black, A. W., Vogel, S., & Woszczyna, M. (2006). Flexible speech translation systems. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(2), 403-411. Retrieved from https://doi.org/10.1109/TSA.2005.860768. Accesed on.

Seeber, K. G. (2007). Thinking outside the cube: Modeling language processing tasks in a multiple resource paradigm. In *Eighth Annual Conference of the International Speech Communication Association* (pp. 1382-1385). Retrieved from http://dx.doi.org/10.21437/Interspeech.2007-21. Accesed on.

Seeber, K. G. (2011). Cognitive load in simultaneous interpreting: Existing theories – new models. *Interpreting, 13*(2). 176–204. Retrieved from https://doi.org/10.1075/intp.13.2.02see. Accesed on.

Seligman, M. (2000). Nine issues in speech translation. *Machine Translation, 15*, 149-186. Retrieved from http://dx.doi.org/10.1023/A:1011180928513. Accesed on.

Singh, P., & Singh, A. (2014). A text to speech (TTS) system with English to Punjabi conversion. arXiv preprint arXiv:1411.3561. Retrieved from https://doi.org/10.48550/arXiv.1411.3561. Accesed on.

Sridhar, V. K. R., Chen, J., Bangalore, S., Ljolje, A., & Chengalvarayan, R. (2013). Segmentation strategies for streaming speech translation. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: *Human Language Technologies* (pp. 230-238).

Straniero Sergio, F., & Falbo, C. (2012). *Breaking ground in corpus-based interpreting studies.* Peter Lang.

Tognini-Bonelli, E. (2001). Corpus linguistics at work. *Corpus Linguistics at Work*, 1-236.

Tripepi Winteringham, S. (2010). The usefulness of ICTs in interpreting practice. *The Interpreters' Newsletter*, *15*, 87–99.

Waibel, A., Badran, A., Black, A. W., Frederking, R., Gates, D., Lavie, A., ... & Zhang, J. (2003). Speechalator: two-way speech-to-speech translation in your hand. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Demonstrations* (pp. 29-30).

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 63–102). Academic Press.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, *3*(2). 159–177. Retrieved from https://psycnet.apa.org/doi/10.1080/14639220210123806. Accesed on.

Williams, J., & Chesterman, A. (2002). *A beginner's guide to doing research in translation studies*. Jerome Publishing.

محمدی، ع. م.(۱۴۰۱) . تحلیلی بر راهبردهای مترجم شفاهی همزمان ایرانی بر اساس نظریه معادل های ترجمه مطالعه گفتمان نماهای استنباطی و توالی. *پژوهشهای زبانشناختی در زبانهای خارجی*، ۱۲ (۱)، ۱۵۲–۱۳۲.
https://doi.org/10.22059/jflr.2021.329419.880.

ولی‌پور، ع. (۱۴۰۰) .داده های بنیادین استراتژیک زبان شناسی، کلید اکمال متقابل تکنولوژی و هوش مصنوعی در حیطه ترجمه ماشینی. *پژوهشهای زبانشناختی در زبانهای خارجی*، ۱۱ (۳)، ۵۵۱–۵۴۱.
https://doi.org/10.22059/jflr.2021.331469.900.