Effect of Automatic speech recognition on translation quality of automatic subtitling platforms: Experimental study of "Veed" and "Iflyrec"

Abstract: The advancements in neural machine translation technology and the usage of speech recognition technology are opening up new possibilities for global online automatic subtitling platforms. A growing number of researchers acknowledge the significant problem of error propagation from automatic speech recognition (ASR) to machine translation (MT) within the prevailing cascade approach to speech translation. But to what extent can ASR influence MT in automatic subtitling translation? To this end, this research conducts a comparative experiment using "Veed" from the UK and "Iflyrec" from China. By assessing the automatic subtitling translation quality of "Veed" and "Iflyrec" under both the original recognition ability and completely accurate recognition conditions, our study demonstrates that their automatic subtitling quality can be improved by 53 times and 28 times respectively, when their recognition ability increased from the initial accuracy level of about 87% to 100%. Our research also shows that in addition to ASR accuracy, the translation quality of automatic subtitling platforms is also related to the translation engine used.

Keywords: Online automatic subtitling platforms; ASR; "Veed"; "Iflyrec"; MT

I. Introduction

Recently, subtitling has become a must-have for video makers for a variety of reasons. With the growth of websites and streaming platforms internationally in recent years, the amount of audiovisual content available online has dramatically increased. As a means of drawing in foreign viewers, more emphasis has been placed on subtitling, and the need for video subtitling is growing. Besides, societies worldwide have become increasingly aware of accessibility requirements for users with a range of disabilities, especially Subtitling for the Deaf and hard of Hearing (SDH) (Bain et al., 2005). Scholars have already found that annotating online videos increases deaf persons' access to the Internet, even when the captions are generated automatically (Shiver & Rosalee, 2015). Meanwhile, people's watching habits also play an important role in the video transmission pattern (Álvarez et al., 2016), adding captions increases the flexibility that videos may be watched on multiple platforms. A study conducted by Verizon Media and Publics Media in 2019 already found that 69% of audiences view videos with sound off in public places, and 80% of consumers are more likely to watch an entire video when captions are available, making video captions critical. All these phenomena have also led to a huge demand for subtitles that are becoming more and more difficult to satisfy only with human resources.

Facing the ever-increasing demand for subtitle translation, how to successfully translate a large number of film and television works within a limited time has become an urgent problem to be solved. However, adding captions and subtitles in videos can be cumbersome and time-consuming as the proper syncing of the subtitles is mandatory both for the audio and the video (Castro et al., 2022). New machine solutions have emerged as a method to deal with the explosion of digital content that needs subtitles and captions (Romero-Fresco & Pérez, 2015). The efficiency of subtitle translation has increased, and the use of automatic subtitling has gained popularity due to the improved ASR or Artificial Intelligence (AI) transcription and Neural Machine Translation (NMT). Many automatic online automatic subtitling platforms such as "TransWAI", "Iflyrec", "Veed", and "Kapwing" have appeared in the global markets.

Automated subtitling is a cutting-edge technical translation practice that creates new working methods for industry experts and regular individuals with video translation needs. The advantage of automatic translation platforms lies in that they present a complex workflow involving many different online tools designed to perform specific tasks such as: automatic transcription, machine translation, and automatic subtitling, which greatly simplifies the video translation procedure compared to the traditional online automatic subtitling platforms (Karakanta et al., 2022). However, like other machine translation engines, their reliability and accuracy are not always guaranteed. Testing their output performance is a method to better understand how these applications function and measure the efficiency of those systems and assess their limitations. (Varga, 2021).

Prior studies have already found several factors that influence MT output quality. Firstly, different language pairs can lead to different MT output quality. According to Shadiev, Sun and Huang (2019), O'Brien et al. (2018), and Ruiz and Federico (2014), translation difficulty increases between distant language pairs. Secondly, the MT results also vary depending on the engine used (Groves & Klaus, 2015). For instance, Google Translate functioned better with Western languages (Aiken & Shilpa, 2011). This highlights the need for careful selection of the translation engine to be used for different translation tasks. Thirdly, the quality of the source text also determines the quality of MT outputs to some extent, such as: length, text difficulty, lexical and structural complexity, and punctuation (Clifford et al., 2013; Jolley and Maimone, 2015; Shadiev, Sun & Huang, 2019). Unlike human translators who have the ability to understand and interpret the context of the translation, machine translation systems only rely on the information provided in the input text. Therefore, if the input material is ambiguous, lacks information, or contains errors, the quality of the machine translation output would be negatively affected.

In order to analyze the effect of ASR ability on automatic translation quality, and provide companies and websites with feedback for further improvements, special attention is paid to the relation between ASR ability and output MT performance, the relation between the ASR ability and platform language background, and the relation between the output MT performance and MT engine. With "Veed" and "Iflyrec" serving as two exemplary platforms, this paper addresses the following questions:

1) Does "Veed" have higher ASR accuracy on English-spoken video material due to English being its native language?
2) Does "Veed" have higher automatic subtitling translation quality if it already has higher ASR accuracy?
3) Does the MT engine also affect the automatic subtitling translation quality?

This paper starts by examining the current advancements in audiovisual translation (AVT) technologies and automated translation platforms, as well as analyzing related research on automated subtitling. Using "Veed" and "Iflyrec" as two representatives, this paper then evaluates their output performance of ASR capability and automatic subtitling translation accuracy by utilizing WER and BLEU metrics. Our goal is twofold: to provide readers with specific examples to better understand the current translation capabilities of online automatic subtitling platforms in the global market and to confirm the integral role of ASR in the quality of automatic subtitling. Additionally, this paper aims to investigate whether the ASR capability and automatic subtitling quality are linked to the platform's language background. For instance, recognition accuracy for English videos may be higher on foreign platforms due to their habitual use of the language; however, translation quality may not be as proficient as that found on Chinese platforms. If this holds true, to what extent does the discrepancy exist, and what possible enhancements could be implemented for both platforms?

II. Recent studies on automatic subtitling translation quality

Subtitling has become an essential tool for the global distribution of audiovisual content, allowing viewers to understand the dialogue of foreign-language films, television shows, and other video formats. However, manual subtitling is a time-consuming and costly process that may limit the availability of content to non-native speaker audiences. Online automatic subtitling, also known as computer-assisted or machine translation subtitling, has emerged as an alternative solution to this challenge. Automatic subtitling has recently received growing interest as MT systems are more frequently used by professionals and companies in various scenarios that provide linguistic solutions and allow users to avoid linguistic barriers. AVT provides a worldwide link to connect people and cultures and different cultures are successfully transmitted through audio-visual products with the aid of AVT by reducing cultural barriers to promoting understanding (Tee et al., 2022).

Prior studies in this field have explored various topics, including the didactic value of automatic subtitling, viewer opinions and receptiveness towards automatic subtitling with a special focus on the perspectives of individuals who rely on subtitles for access (e.g., SDH viewers), as well as the effectiveness, applicability, and precision of automatic subtitling for different platforms.

Some researchers have explored the didactic value of automatic subtitling. Since automatic subtitles can greatly enhance the accessibility, comprehension, and retention of educational materials, automatic subtitles have also been increasingly used in the field of education, particularly in assisting simultaneous interpretation, vocabulary learning, and teaching. Simultaneous interpretation is a particularly challenging task for many individuals, and the use of automatic subtitles provides a visual cue that can make it easier for individuals to follow what is being said in real time. A study by Visky (2015) shows that subtitling is an effective method for improving students' interpretation skills, and the students who receive subtitling instruction perform significantly better than the control group who do not receive this instruction. In terms of vocabulary learning, automatic subtitles can provide contextualized learning opportunities. Research indicates that the use of automatic subtitles can significantly improve vocabulary learning among language learners (Montero Perez et al., 2015). Moreover, in teaching, automatic subtitles can provide support for students with different learning needs. A study conducted

by Denis Burnham and his colleagues in 2008 finds that the use of automatic subtitles for educational videos positively impacts the engagement and comprehension of deaf and hard-of-hearing individuals.

Besides, audience attitude and feedback toward automatic subtitling especially perspectives from SDH people is another key research topic. AVT has emerged as an interdisciplinary field that encompasses not only subtitling but also other modalities of audiovisual content translation, such as audio description and sign language interpretation, etc. As such, several studies have investigated the applications of automatic subtitling in improving accessibility and inclusion for people with disabilities. For example, Iriarte (2014) reveals that subtitling may help SDH people better understand the content and dialogue of the media they are consuming. Another survey conducted by Butler (2019) among deaf and hard-of-hearing individuals implicates the need for improved caption quality and consistency to better serve this population. Though online automatic subtitling holds significant potential for expanding the accessibility of audiovisual content globally, further research is needed to address the limitations of AI algorithms in handling idiomatic expressions, colloquial language, and punctuation.

In addition, more and more scholars acknowledge the importance and usefulness of subtitles, studies have shown that subtitles and captions are readily attended to by diverse viewer groups; viewers of subtitled media have the ability to process subtitles and captions in an efficient manner and take in the salient elements of the visuals simultaneously (Perego et al., 2010). In view of the difficulty in guaranteeing the quality of automatic subtitles, scholars also conduct relevant research. One survey by Hu, Ke, Sharon O'Brien, and Dorothy Kenny (2020) focuses on the use and utility of MT for MOOC content to test the impact machine-translated subtitles have on Chinese viewers' reception of MOOC content by using an eye-tracking experiment and survey methods. Another survey by Sivakorn Malakul and Innwoo Park (2023) suggests that in the language pair of English-Thai, integrating an auto-subtitle system into MOOC could effectively enhance comprehension, and reduce the cognitive load of Thai secondary school students. For online automatic subtitling platforms. Miller (2023) describes available online tools and methods nowadays, and demonstrates how to translate videos on "Veed", "Kwaping", and "Youtube". Other scholars also mention these platforms in their research (Alabsi, 2020; Kanmounye, 2022; Yildirim, 2023) without delving deeper into their functionalities or qualities.

A review of the literature indicates that there is a dearth of research that specifically investigates the ASR capability and its influence on the quality of automatic subtitling, and less attention has been paid to the thriving automatic translation platforms. This study expands on the existing research by examining the influence of ASR on the translation quality of automatic subtitling for both Chinese and UK translation platforms through objective and subjective evaluations.

III. Methodology

3.1 Research design

The main aim of this study is to assess and compare the ASR ability of online automatic subtitling platforms in China and the UK, to determine whether the ASR ability offers a significant difference to the quality of automatic translation. Choosing suitable platforms, video material, and evaluation metrics is the fundamental step.

(1) Automatic subtitling translation platforms

"Iflyrec" and "Veed" are chosen as two representatives for several reasons as they share some similarities and differences for comparison. Firstly, they share the same potential user group, and they are all convenient and easy to operate for both professionals and normal users with their friendly UI design. Before this study starts, the author nearly finds all the automatic subtitling platforms used in the Chinese domestic market in order to evaluate the overall performance. In view of the fact that some platforms only provide trial opportunities for cooperative enterprises, individuals are not able to use them in this research, so this research does not involve the evaluation of such enterprise-only platforms. As "Veed" states on its website, unlike traditional video editing products, "Veed" is trying to make video editing accessible to all, "Veed" focuses on accessibility, ease of use and is building tools in line with the type of consumption nowadays. Users only need to edit a few words to translate the whole transcript to their preferred language. Secondly, they all exclaim that they have high accuracy scores. According to its website, "Veed" has an incredible 95% accuracy in generating subtitles and translation; while "Iflyrec" says its transcription accuracy is 97.5%. Thirdly, they boast their own unique advantages which make them more popular with customers. "Iflyrec" is an online automatic subtitling platform developed by IFLYTEK, a company famous for its voice recognition technology in the general public. According to the "China Intelligent Speech Transcription Tool Industry Insights 2021"[1] report released by Analysys, a well-known third-party research organization in China, iFLYTEK ranks first in brand awareness, with 90% of the respondents indicating that they know the brand. While "Veed" was founded in London, UK, in 2018, with a team size of over 150 and 14,129 followers on Linkedin. As shown in Figure 1. below, the unique advantage of "Veed" lies in its highlight of "Low Confidence Word" or automatic detection of potential subtitle errors both temporal and linguistic, which is deemed as a technological advance in Audiovisual Translation (Bolaños-García-Escribano et al., 2021).

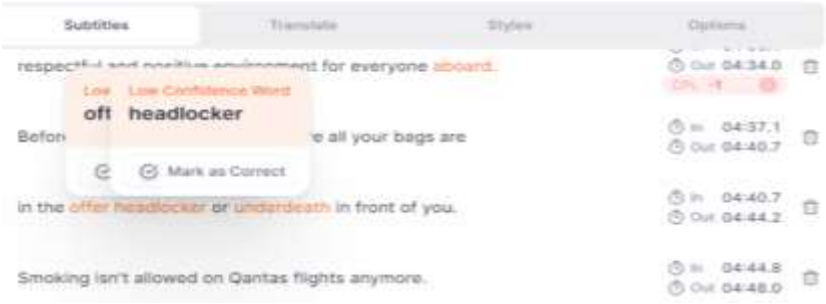

Figure 1. The setting of "Low Confidential word"

(2) Video material: Qantas Safety Video

For the video material, this paper picks Qantas Safety Video[2] 2020 Centenary from Qantas official YouTube account, it is a video created by Qantas Airways to inform passengers about their onboard safety measures in an engaging and approachable way. Its use of celebrities and clear, simple language help to make the safety instructions more accessible and memorable, while its visually engaging and upbeat style helps to hold viewers' attention and keep them engaged throughout the video.

[1] analysys.cn
[2] https://www.youtube.com/watch?v=rLq8if1nkTM

This video has some characteristics such as language understandability, different accents, and the use of terminology. All these make it suitable for serving the purpose of this paper - testing the ASR quality and automatic subtitling of automatic translation platforms. The main features of the video are shown in Table 1 and described below.

Table 1. Features of 'Qantas Safety Video'

Name	Qantas safety video
Language	English
Domain	Aviation
Duration	00:08:15
Lines of subtitle	115
Total words	930
Text difficulty	610 -1000L

In terms of language use, the video is notable for its clear and concise language, which is tailored to the needs and expectations of its audience. This approach helps to ensure that viewers are able to understand and follow the safety procedures outlined in the video, which is ultimately the most important objective of the video. After being tested, the Lexile Rouge of this text is about 610L-1000L[3], main factors such as word frequency and sentence length of this video show that this text is not difficult to comprehend. In addition, while the inclusion of people from diverse backgrounds helps to underscore the airline's commitment to inclusivity and community, the different ages, genders, and accents of the people in the video can also help us to understand and analyze factors that impede high-quality speech recognition. Lastly, one of the key features of the video is its use of aviation terminology, which is of great help in testing the accuracy of the terminology translation because the use of proper terminology is one aspect of translation quality that is especially significant in academic research (GHENŢULESCU, 2015). For audiences, this adds a level of authenticity and authority to the safety briefing, and helps to build passenger trust and confidence. Moreover, the importance of terminology in translation quality is demonstrated by its inclusion in various metrics for assessing translation quality, such as the Translation Quality Evaluation (TQE) and the Common Sense Advisory (CSA) benchmarking program. These metrics take into account the accuracy of terminology translation in evaluating the overall quality of a translated text. Given all these characteristics, this paper selects Qantas Safety Video as the research material.

3.2 Data collection

In this research, 2 experiments are conducted and 5 data sets are collected to test the quality of ASR output and automatic subtitling of "Veed" and "Iflyrec".

In the first experiment of ASR accuracy comparison, 1 set of speech-to-text data (hereafter Veed $_0$ and Iflyrec $_0$) is generated after submitting the "Qantas Safety video" on "Iflyrec" and "Veed". The reference speech-to-text data is transcribed from Qantas' official channel on YouTube.

[3] Lexile Text Analyzer | Lexile & Quantile Hub

In the second experiment of automatic subtitling translation comparison, 3 sets of data are collected and analyzed. The author uses the original speech-to-text data of "Iflyrec" and "Veed" to generate one set of automatic subtitling translations (hereafter Veed $_1$ and Iflyrec $_1$), and then inputs the speech-to-text data on these two platforms to generate the final automatic subtitling translation version (hereafter Veed $_2$ and Iflyrec $_2$). For the human translation reference, this paper uses the translation from a tutor affiliated with the Civil Aviation University of China.

3.3 Date evaluation

In translation studies, the issue of translation quality has always been of great importance. As recent developments in machine translation and speech translation are opening up opportunities for computer-assisted translation tools with extended automation functions, automatic scoring systems are also used for the evaluation of ASR ability and machine translation. In order to eliminate the effect of subjectivity on translation quality assessment (TQA), assessment should be done based on predefined criteria and models (Foradi et al., 2022). However, studies also show that automatic evaluation metrics are not always reliable and may suffer from low correlation with human evaluations, and human evaluation is essential to determine the accuracy and naturalness of the MT system's output, the use of both automatic evaluation and manual evaluation helped to identify errors and improve the overall accuracy and quality of the system's output (Chatzikoumi, 2020; Rivera-Trigueros, 2022). Therefore, this paper combines automatic evaluation and human evaluation to obtain accurate and reliable results and to identify errors and inconsistencies that may not be caught by automatic evaluation alone. The following part introduces the evaluation metrics and tools of ASR output and automatic subtitling translation respectively.

(1) For ASR evaluation

WER is an important evaluation metric for ASR which is used to evaluate the accuracy of automatic speech transcription systems or how accurately the ASR system transcribes spoken words into text. WER was first proposed in the 1970s as a metric for evaluating speech recognition systems (Hirsch & Pearce, 2000). WER is defined as the proportion of incorrectly recognized words in the total number of words spoken by the user. In other words, WER calculates the number of substitution, insertion, and deletion errors in the transcription output (Young & Chase, 1998).

The calculation of WER is typically done using the following formula: WER = (S + I + D) / N, lower WER often indicates that the ASR software is more accurate in recognizing speech. A higher WER, then, often indicates lower ASR accuracy.

The author inputs Veed $_0$ and Iflyrec $_0$ into an online WER test tool[4] "Amberscript" to test the WER score. As stated on its website, factors that can affect WER, without necessarily reflecting the capabilities of the ASR technology itself, include microphone quality, speaker pronunciation, background noise, unusual names, and so on. The online tool used here can only calculate the WER scores, and can't give a specific error analysis to further

[4] https://www.amberscript.com/en/wer-tool/

explain what aspects of these platforms should be improved, the author manually classified the error types into substitution error, insertion error and deletion error based on the WER's calculation formula. Human evaluation and analysis are conducted.

(2) Automatic subtitling evaluation

Machine translation evaluation models are used to assess the quality of machine-generated translations. Generally speaking, the metrics for evaluating the quality of MT have relied on assessing the similarity between an MT-generated hypothesis and a human-generated reference translation in the target language. There are different evaluation metrics, such as BLEU, or Bilingual Evaluation Understudy, TER and METEOR, but the most commonly used one is BLEU. The BLEU evaluation method was first proposed by Kishore Papineni and colleagues in 2002 as a metric for machine translation evaluation and its calculated formula is shown in figure 2 below. The BLEU score measures the similarity between the machine-generated output and the reference translations based on n-grams, which are contiguous sequences of words. The score ranges from 0 to 1, with 1 indicating a perfect match between the machine-generated output and the reference translations (Kishore Papineni et al 2002). A score of 0.6 or 0.7 is considered the best one can achieve (Štajner et al., 2015).

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$$

Then,

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

Figure 2. Bleu Score formula

BLEU scores have been used to evaluate various machine translation systems, including rule-based, statistical, and neural machine translation systems, and thus it has become a standard metric in the field of machine translation evaluation (Koehn, 2004). Compared to other translation quality metrics such as TER (translation error rate) and METEOR (metric for evaluation of translation with explicit or obscured reference), BLEU is more efficient in measuring syntactic correctness and fluency, as well as word-level accuracy (Coughlin, 2003; Shiver & Rosalee, 2006). BLEU is also able to handle multiple reference translations, which makes it more robust than other metrics (Banerjee & Lavie, 2005). Based on these factors, this paper selects BLEU as the automatic translation quality evaluation metric and uses an online tool[5] "Lestmt" to get the BLEU score automatically. Considering the reliability and correlation with human evaluations, this paper uses the core typology of Multidimensional Quality Metrics (MQM 2.0)[6], the most recent iteration of the framework to measure the quality of automatic subtitling translation generated by "Iflyrec" and "Veed". According to the official website, MQM is a framework for analytic TQE. It can be applied to both human translation and machine translation, which makes it suitable for this research purpose. The seven high-level error type dimensions of MQM is: terminology, accuracy, linguistic conventions, style, locale conventions, audience appropriateness, design and markup.

[5] https://www.letsmt.eu/Bleu.aspx
[6] What Is MQM?

IV. Results and analysis

The results sections are subdivided into two parts: ASR quality evaluation and automatic subtitling evaluation. As mentioned before, this paper selects "Iflyrec" and "Veed" as two representatives to test their quality of ASR and automatic translation subtitling. After uploading the video material separately on these two platforms, one set of speech-to-text data (Veed $_0$ and Iflyrec $_0$) and two sets of automatic translation subtitling (Veed $_1$ and Iflyrec $_1$, Veed $_2$ and Iflyrec $_2$) are generated and then tested individually with reference by online tools and human assessment.

4.1 ASR evaluation result for "Iflyrec" and "Veed"

After getting Veed $_0$ and Iflyrec $_0$ automatically recognized and generated by "Iflyrec" and "Veed", the author uploads them on the "Amberscript" website to automatically calculate the WER score. WER scores and their main features are recorded and summarized in table 2 below.

Table 2. WER score

	Iflyrec $_0$	Veed $_0$	Reference
Lines of subtitle	173	125	115
Total Words	907	906	896
WER score	0.129	0.115	N/A

As shown in Table 2, there are in total 115 lines of subtitles and 896 words in the subtitling of the original English subtitling, while both "Ilfyrec" and "Veed" fail to fully recognize them all. The WER score of "Iflyrec" is a little bit higher than that of "Veed", which indicates that "Veed" has better ASR quality and performs better in speech-to-text in English-spoken material.

Since "Amberscript" can only calculate the WER score without any explanation about the specific error analysis to further explain what aspects of these platforms should be improved, the author manually classified the error types into substitution error (S), insertion error (I) and deletion error (D) based on the WER's calculation formula. Substitution errors include misspellings, proper nouns, and countable and uncountable nouns, etc. Errors of insertion occur when the auto-generated captions fail to record a word or phrase that the speaker said, while errors of deletion occur when certain words or phrases appear in the captions that were never spoken in the video. These three classifications cover every error that occurs in the subtitling and is manually assessed by the author.

In order to simplify the counting procedure, each line of subtitling that appeared on the screen is regarded as a counting unit when doing the human assessment. After annotating the error type of Veed $_0$ and Iflyrec $_0$, it is found that there are in total 23 errors in Iflyrec $_0$ and 19 errors in Veed $_0$. For "Iflyrec", 4 of these 23 errors is insertion error and 19 of them is substitution error; while for "Veed", 18 of 19 errors is substitution error, and 1 deletion error. Some examples are shown in Table 3.

Table 3. ASR error category

Human reference	Iflyrec		Veed	
	Iflyrec $_0$	Error type	Veed $_0$	Error type
Longreach	language	S	Longrich	S
flashier	pleasure	S	closer	S
Keep it done up low and **tight**	keep it done up low and typed	S	Keep it done up low and tight	/
pass the strap around your waist	How's the strap around your waist	S	Pass the strap around your waist	/
hold on to your lower **legs**	hold on to your lower legs	/	hold on to your lower lens.	S
Lights will guide you to your exits	lights will guide you to your exits.	/	like this guide you to your excess.	S
in the **overhead** locker	in the overhead locker	/	in the offer head locker	S
smoke **detectors**	smoke detectives	S	smoke detectors	/
And today, with your help	Today with your help	I	and today, with your help	/
operate **in** an emergency	operate an emergency	I	operate in an emergency room	D
help children **in** need	help children need	I	help children leave	S
Whether you are starting your journey or heading home	you start your journey or heading home	I	Whether you are starting your journey or heading home	/

It can be seen that both "Iflyrec" and "Veed" are able to transcript the majority of the spoken words, but there are still many mistakes and errors, such as misrecognized words and misplaced syntax. In this research, the main error type in English-spoken material transcription data is substitution error, which shows these platforms fail to recognize some of the expressions and information in the original video owing to various reasons such as misidentification of terms, accent, loud background music and speaking speed, etc. Our test results show that while "Iflyrec" and "Veed" are capable platforms for generating ASR and subtitling translations to assist humans in the transcription process, their accuracy rates are not yet high enough to produce high-quality translations without human supervision or manual correction. For higher ASR accuracy, video noise reduction, speech

annotation, and pre-training on different terms, accents, and language grammar structure analysis are needed for these platform providers especially when dealing with complex and highly professional topics.

4.2 Automatic subtitling evaluation result for "Iflyrec" and "Veed"

In this part, the author uses Veed $_0$ and Iflyrec $_0$ to generate one set of automatic subtitling Veed $_1$ and Iflyrec $_1$ and uses accurate source English subtitling to generate another set of automatic subtitling Veed $_2$ and Iflyrec $_2$. BLEU scores of these two sets are automatically calculated by the "letsmt" platform and summarised below.

Table 4 BLEU score

	Lines of subtitle	**Total Words**	**BLEU Score**
Iflyrec $_1$	171	1462	0.0044
Veed $_1$	125	1716	0.0019
Iflyrec $_2$	115	1427	0.1244
Veed $_2$	115	1597	0.1013
Human reference	115	1359	/

According to Table 3 and Table 4, two characteristics can be found:

Firstly, as it can be seen from the data, with a WER score of 0.0015, "Veed" has more accurate ASR quality than "Iflyrec", while the BLEU score of Veed $_1$ is slightly lower than Iflyrec $_1$, which indicates that its automatic subtitling translation is not as good as that of "Iflyrec" even it has more accurate source material. This can be further supported by the BLEU score comparison between Veed $_2$ and Iflyrec $_2$. This finding answers question 2 proposed in the previous chapter that despite using the same source material or more precise source material, "Veed" underperforms "Iflyrec" in English to Chinese subtitling translation.

Secondly, at least for "Veed" and "Iflyrec", their engines or algorithms tend to have different habits for processing video materials. On one hand, both "Iflyrec" and "Veed" tends to follow the original subtitle structure without obvious combination or segmentation of the source subtitling when handling the automatic subtitling translation tasks online. From the table above, there are 173 lines of subtitles in Iflyrec $_0$ and 171 lines of subtitles in Iflyrec $_1$, while in Veed, there are 125 lines of subtitles in Veed $_0$ and 125 lines of subtitles in Veed $_1$. The line of source subtitles is almost the same as the line of automatic subtitles. On the other hand, the engine used by "Iflyrec" tends to segment the video material more frequently, while the engine used by "Veed" is prone to containing more words in automatic subtitling. If we compare the features of Iflyrec $_0$ and Veed $_0$ with the human reference, there are 173 lines of subtitles in Iflyrec $_0$ and 125 lines of subtitles in Veed $_0$, which all exceed 115, the line of the subtitles of human reference. This shows that "Iflyrec" tends to segment or cut the subtitle of the video material more frequently when processing ASR tasks. In the comparison of the subtitles' total words, there are 1359 words in human reference, 1462 words in Veed $_1$, and 1427 words in Veed $_2$, while there are 1716 words in Veed $_1$ and 1597 words in Veed $_2$. This shows that the translation engine used by "Veed" is prone to containing more words

when doing automatic subtitling.

Based on these analyses, it is thus assumed that "Veed" has a more accurate ASR ability than that "Iflyrec" in the English-spoken material, which supports our first question. Meanwhile, the result also shows that "Veed" needs to improve its automatic translation quality given its lower subtitling translation score under more accurate speech-to-text ability. Considering the unreliability of machine assessment and in order to find more specific error types for further improvements, this paper uses the MQM core typology to evaluate their automatic subtitling translation performance. Similarly, each line of subtitling is regarded as a counting unit when doing the human assessment for simplifying the evaluation procedure. After annotating the error type of the final automatic subtitling translation data of "Iflyrec" and "Veed", the error type summary is gained in Figure 2 below.

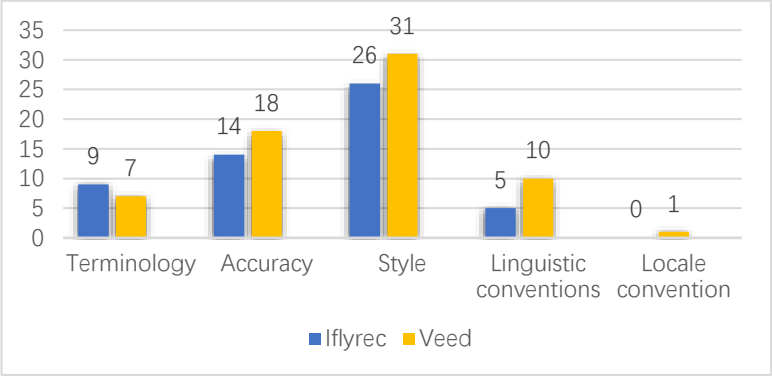

Figure 2. Error summary of automatic subtitling translation

In general, both "Iflyrec" and "Veed" contain a certain degree of errors in their automatic subtitling translations, with "Veed" gaining a higher error rate than "Iflyrec". The predominant error type for both systems is related to style error, followed by accuracy error and terminology error. Moreover, the number of error lines in the subtitling of "Veed" is higher than that of "Iflyrec", indicating the need for improvement on these aspects in the En-CH translation. This is important because inaccurate or erroneous subtitling translations can lead to misinterpretation of the actual content and negatively impact the user experience. Following are some error examples from "Veed" and "Iflyrec".

One of the most apparent mistakes is related to style, which is evident in the following examples. The reference translation contains the word "您" (a courteous translation version of You) 39 times and the word "请" (a courteous translation of Please) 15 times. However, "您" is only used 12 times in "Iflyrec" and only 5 times in "Veed", and "请" is used 7 and 5 times, respectively. It should be noticed that the video selected for this research is an aviation safety instruction for passengers onboard and the audience. Airline belongs to the service industry, and its language naturally has the characteristics of politeness and professionalism, which also needs to be conveyed in the target translation.

Besides, terminology error is another key error type. For instance, the phrase "brace position" in the aviation industry should be translated into "防撞击姿势" (an instruction that can be given to prepare for a crash, such as on an aircraft) instead of "支撑架的位置", which means the position of the brace; "overhead locker" should be

translated into “上方行李架” instead of “头顶的储物柜”.

Apart from these, accuracy error is another error type that deserves to be noticed. If the automatic subtitling fails to understand the meaning of words or sentences in the context of the source text, misunderstanding occurs as the same words may have different meanings in a different context. For example, during a demonstration of how to wear an oxygen mask, the video says, “Put it on quickly and tighten the strap” (which means “拉紧带子”), this phrase is translated into “拧紧带子” which is not a common and logical expression in Chinese. In another subtitling, “If you can reach the seat in front, brace by folding your arms”, the word “reach” is translated into “到达” by “Veed”, which is not appropriate in this context.

Overall, the MQM core typology evaluation supports the previous findings that “Veed” needs to improve its automatic translation quality, specifically in terminology, style, and translation accuracy despite having a higher ASR ability.

V. Conclusion

Translation platforms, particularly those that offer automatic transcription and translation capabilities, have become indispensable tools for global communication. As these technologies become increasingly advanced, it is becoming more important to assess their ASR capabilities and automatic subtitling translation quality. This paper selects “Iflyrec” and “Veed” as two representatives to test the ASR ability and automatic subtitling translation quality in the EN-CH direction by both automatic and manual approaches. Following conclusions may be drawn from the experimental results.

Firstly, the improvement of speech recognition ability can improve the quality of translation. As demonstrated in the previous chapter, “Iflyrec”’s BLEU score increased from 0.0044 to 0.1244, or 28 times higher, when its ASR accuracy increased from 87% to 100%. Similarly, “Veed”’s BLEU score improved from 0.0019 to 0.1013, or 53 times higher, when its ASR accuracy increased from 88.5% to 100%. These findings suggest that language recognition plays a critical role in automatic translation platforms, and optimizing speech recognition can significantly enhance translation quality.

Secondly, both ASR capability and automatic subtitling translation accuracy are affected by the platform language background to some extent. As can be seen from the WER score of ASR comparison, “Veed”, with English as its habitual language, has a higher WER score than “Iflyrec”. A comparison of the WER scores of “Veed” and “Iflyrec” in ASR reveals that “Veed”’s automatic subtitling translation scores are consistently lower than “Iflyrec”, regardless of whether “Veed” has the same or even higher ASR ability. This indicates that “Veed” should improve its MT algorithm, data training, especially in the EN-CH direction, while “Iflyrec” needs to improve the recognition ability of other languages.

Thirdly, the accuracy of automatic subtitles cannot be guaranteed without manual proofreading. Although the ASR ability of the current platform is relatively precise, the quality of machine-generated subtitles is low and cannot be used directly as seen from the BLEU score table above. Even when the platform can fully recognize the text, as demonstrated by the second experiment in this paper, the BLEU score remains very low. Upon applying the MQM metric to analyze the errors in the automatic translation subtitling, it can be seen that both platforms have terms translation errors, style errors, mistranslations, etc., indicating further optimization of the machine

translation algorithm is necessary for both platforms. For the current automatic subtitle translation, human intervention and proofreading are essential and necessary to ensure semantic accuracy and optimize the subtitles.

The increasing availability and operability of AVT software, datasets, and communities are leading to more opportunities for collaboration between language technology and AVT and offering a more positive and promising time for industry and academia. Our test results show that while "Iflyrec" and "Veed" are capable platforms for generating ASR and subtitling translations, their output performances are not yet accurate enough to produce high-quality transcriptions and translations without human supervision or manual correction. It is imperative for both "Iflyrec" and "Veed" to continue refining their automatic subtitling algorithms to achieve higher levels of accuracy and precision in their translations.

Although our studies can contribute to a better understanding and utilization of automatic subtitling technologies, there are many issues that deserve to be investigated to facilitate practical applications in various domains, such as education, entertainment, or accessibility. For instance, in terms of material selection and research platforms, this paper only takes "Veed" and "Iflyrec" as two representatives, one aviation video as an example to analyze the output performance of current online automatic subtitling platforms, future research on automatic subtitling translation can be extended into different platforms and material genres. Meanwhile, this paper only analysis two languages: English and Chinese, and as MT is constantly evolving and online automatic subtitling platforms support other languages, more updated research on MT accuracy in various language pairs also becomes essential, Future research on these issues will provide more insight into the output performance and the utilization of online automatic subtitling platforms.

References:

[1] Aiken, M., & Balan, S. (2011). An analysis of Google Translate accuracy. *Translation journal*, *16*(2), 1-3.

[2] Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., ... & Del Pozo, A. (2016). Automating live and batch subtitling of multimedia contents for several European languages. *Multimedia Tools and Applications*, 75(18), 10823-10853.

[3] Alabsi, T. (2020). Effects of adding subtitles to video via apps on developing EFL students' listening comprehension. *Theory and Practice in Language Studies*, 10(10), 1191-1199.

[4] Bain, K., Basson, S., Faisman, A., & Kanevsky, D. (2005). Accessibility, transcription, and access everywhere. *IBM systems journal*, 44(3), 589-603.

[5] Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65-72).

[6] Bolaños-García-Escribano, A., Díaz-Cintas, J., & Massidda, S. (2021). Latest advancements in audiovisual translation education. *The interpreter and translator trainer*, 15(1), 1-12.

[7] Burnham, D., Leigh, G., Noble, W., Jones, C., Tyler, M., Grebennikov, L., & Varley, A. (2008). Parameters in television captioning for deaf and hard-of-hearing adults: Effects of caption rate versus text reduction on comprehension. *Journal of deaf studies and deaf education*, 13(3), 391-404.

[8] Butler, J. (2019). Perspectives of deaf and hard of hearing viewers of captions. *American Annals of the Deaf*, 163(5), 534-553.

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

[10] Clifford, J., Merschel, L., & Munné, J. (2013). Surveying the landscape: What is the role of machine translation in language learning? *The Acquisition of Second Languages and Innovative Pedagogies*, (10), 108-121.

[11] Coughlin, D. (2003). Correlating automated and human assessments of machine translation quality. *In Proceedings of Machine Translation Summit IX,* 63-70.

[12] De Castro, M., Carrero, D., Puente, L., & Ruiz, B. (2011). Real-time subtitle synchronization in live television programs. *In 2011 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)* (pp. 1-6). IEEE.

[13] Foradi, Z., Faroughi, J., & Rezaeian Delouei, M. R. (2022). Assessing the Performance Quality of Google Translate in Translating English and Persian Newspaper Texts Based on the MQM-DQF Model. *Journal of Language and Translation*, 12(4), 107-118.

[14] GHENŢULESCU, L. R. (2015). The Importance of Terminology for Translation Studies. In the Beginning Was the Word. *On the Linguistic Matter of Which the World Is Built. Bucureşti: Ars Docendi*, 54-61.

[15] Groves, M., & Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes*, 37, 112-121.

[16] Hirsch, H. G., & Pearce, D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *In ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW)*, 29-32.

[17] Hu, K., O'Brien, S., & Kenny, D. (2020). A reception study of machine translated subtitles for MOOCs. *Perspectives*, 28(4), 521-538.

[18] Iriarte, M. M. (2014). The reception of subtitling by the deaf and hard of hearing. Preliminary findings. *Translation Research Project*, 5, 63.

[19] Jolley, J. R., & Maimone, L. (2015). Free online machine translation: Use and perceptions by Spanish students and instructors. *Learn languages, explore cultures, transform lives*, 181-200.

[20] Karakanta, A., Bentivogli, L., Cettolo, M., Negri, M., & Turchi, M. (2022). Post-editing in Automatic Subtitling: A Subtitlers' Perspective. *In Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 259-268). European Association for Machine Translation.

[21] Kanmounye, U. S., Mbaye, M., Phusoongnern, W., Moreanu, M. S., Niquen-Jimenez, M., & Rosseau, G. (2022). Neurosurgical training in LMIC: opportunities and challenges. *Learning and Career Development in Neurosurgery: Values-Based Medical Education*, 219-227.

[22] Koehn, P. (2004). Statistical significance tests for machine translation evaluation. I*n Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 388-395).

[23] Malakul, S., & Park, I. (2023). The effects of using an auto-subtitle system in educational videos to facilitate

learning for secondary school students: learning comprehension, cognitive load, and satisfaction. *Smart Learning Environments*, 10(1), 4.

[24] Miller, M. D. (2023). Translation Tools and Techniques. *In Discovering Hidden Gems in Foreign Languages* (pp. 173-225). Cham: Springer International Publishing.

[25] Montero Perez, M., Peters, E., & Desmet, P. (2015). Enhancing vocabulary learning through captioned Video: An eye-tracking study. *The Modern Language Journal*, 99(2), 308-328.

[26] Läubli, S., Fishel, M., Massey, G., Ehrensberger-Dow, M., Volk, M., O'Brien, S., ... & Specia, L. (2013). *Assessing post-editing efficiency in a realistic translation environment*, 237-262.

[27] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

[28] Perego, E., Del Missier, F., Porta, M., & Mosconi, M. (2010). The cognitive effectiveness of subtitle processing. *Media psychology*, 13(3), 243-272.

[29] Rivera-Trigueros, I. (2022). Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, 56(2), 593-619.

[30] Romero-Fresco, P., & Pérez, J. M. (2015). Accuracy rate in live subtitling: The NER model. *Audiovisual translation in a global context: Mapping an ever-changing landscape*, 28-50.

[31] Ruiz, N., & Federico, M. (2014). Complexity of spoken versus written language for machine translation. *In Proceedings of the 17th Annual conference of the European Association for Machine Translation* (pp. 173-180).

[32] Shadiev, R., Sun, A., & Huang, Y. M. (2019). A study of the facilitation of cross-cultural understanding and intercultural sensitivity using speech-enabled language translation technology. *British Journal of Educational Technology*, 50(3), 1415-1433.

[33] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers* (pp. 223-231).

[34] Shiver, B. N., & Wolfe, R. J. (2015). Evaluating alternatives for better deaf accessibility to selected web-based multimedia. *In Proceedings of the 17th international ACM SIGACCESS conference on computers & accessibility* (pp. 231-238).

[35] Štajner, S., Béchara, H., & Saggion, H. (2015). A deeper exploration of the standard PB-SMT approach to text simplification and its evaluation. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 823-828).

[36] Tee, Y. H., Amini, M., Siau, C. S., & Amirdabbaghian, A. (2022). English to Chinese fansub translation of humour in The Marvellous Mrs. Maisel. *Texto Livre: Linguagem e Tecnologia*, 15, 1-21.

[37] Varga, C. (2021). Online Automatic Subtitling Platforms and Machine Translation. *Buletinul Stiintific al Universitatii Politehnica din Timisoara, Seria Limbi Moderne*, (20), 37-49.

[38] Visky, M. (2015). The Use of Subtitling in Teaching Professional Interpretation. *Procedia-Social and Behavioral Sciences*, 191, 2641-2644.

[39] YILDIRIM, S. (2023). Alan eğitiminde Web 2.0 uygulamalarının coğrafya dersi bağlamında değerlendirilmesi. *lnternational Journal of Geography and Geography Education*, (49), 41-58.

[40] Young, S. J., & Chase, L. L. (1998). Speech recognition evaluation: a review of the US CSR and LVCSR programmes. *Computer Speech & Language*, 12(4), 263-279.