Machine Translation Implementation in Automatic Subtitling from a Subtitlers' Perspective

Bina Xie 21439095@life.hkbu.edu.hk
Department of Translation, Interpreting and Intercultural Studies, Hong Kong Baptist University, Hong Kong, China

Abstract

In recent years, automatic subtitling has gained considerable scholarly attention. Implementing machine translation in subtitling editors faces challenges, being a primary process in automatic subtitling. Therefore, there is still a significant research gap when it comes to machine translation implementation in automatic subtitling. This project compared different levels of non-verbal input videos from English to Chinese Simplified to examine post-editing efforts in automatic subtitling. The research collected the following data: process logs, which records the total time spent on the subtitles, keystrokes, and user experience questionnaire (UEQ). 12 subtitlers from a translation agency in Mainland China were invited to complete the task. The results show that there are no significant differences between videos with low and high levels of non-verbal input in terms of time spent. Furthermore, the subtitlers spent more effort on revising spotting and segmentation than translation when they post-edited texts with a high level of non-verbal input. While a majority of subtitlers show a positive attitude towards the application of machine translation, their apprehension lies in the potential overreliance on its usage.

1. Introduction

1.1. Automatic subtitling and machine translation implementation

The progress of technological advancements has led to the expansion of automation in subtitling, transitioning from machine translation (MT) to fully automatic subtitling. Automatic subtitling involves a complex workflow, including auto-transcription, automatic segmentation, auto-spotting and MT. Recently, the audiovisual industry has shown increasing interest in automatic subtitling. Prominent streaming platforms like YouTube and Bilibili have already adopted automatic subtitling. Moreover, several advanced subtitling platforms or software now incorporate automated tools to improve productivity.

Researchers also start to explore experimental research in MT and automatic subtitling. Georgakopoulou (2021) discusses machine translation implementation issues and future trends in MT research, such as intelligent text segmenters, MT quality estimation, and metadata usage. VARGA (2021) analysed machine translation quality from different online automatic subtitling platforms in audiovisual translation (AVT). Inconsistencies were reported, including literal translation, word order, language register, noun-adjective agreement, punctuation, and mistranslation. Karakanta (2022) introduces experimental methods from MT in subtitles to automatic subtitling and points out that automatic subtitling poses extra challenges for MT, such as segmentation and time stamps. Other research focuses on subtitler feedback. Karakanta et al. (2022b) collected subtitle post-editing data to investigate how subtitlers interact with automatic subtitling, through process logs, keystrokes and questionnaire. Karakanta et al. (2022a) analyse feedback from subtitlers on the use of automatic subtitling. Most subtitlers show a positive

Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, pages 54–64
September 4–8, 2023, Macau SAR, China.

attitude towards automatic subtitling. Besides, automatic subtitling helps subtitlers save time and effort on the tedious part of the work. In the end, they call for more automatic subtitling tests by the actual users and sufficient consideration of translators' views. Therefore, research in automatic subtitling still represents a significant research gap.

1.2. Non-verbal input of subtitle translation

Researchers have noticed non-verbal information in audiovisual translation for the last decade. Guillot (2018) proposes that non-verbal information in audiovisual materials is a unique feature of audiovisual translation because translators need to interact visual footage and audio tracks with written subtitles. This kind of information can affect the meaning of films and TV programs (e.g. Perego, 2009) and the decision-making process of subtitlers (e.g. Pérez-González, 2014). Díaz-Cintas and Remael (2014) discover that redundancy between "look, gestures, facial expressions and language" requires extra attention from translators. They expand non-verbal information in AVT that viewer obtains dialogue information from the images rather than from the verbal text, such as pronouns in audiovisual texts. They also introduce semiotic cohesion, the criteria to distinguish texts with different levels of non-verbal input, which are the interaction between images and words and the interaction between gestures and speech. Based on this theory proposed by Díaz-Cintas and Remael (2014), Huang and Wang (2022) compared low level of non-verbal input with high level one in two audiovisual texts. They use eye tracking and keystroke logging to compare post-editing and translation efforts from translation students. The results show that although non-verbal input affected post-editing effort, a higher level of non-verbal input required lower cognitive effort. Therefore, they conclude that the multimodal nature of audiovisual texts may not be an obstacle during the subtitle post-editing process since the texts with more non-verbal input are likely to help the translators.

Based on Huang and Wang's research, this study adopts a mixed method, combining process logs, keystrokes, and questionnaire to compare subtitlers' post-editing efforts. Unlike previous research, which has been focused on machine-translating human-generated source language subtitles, the experiment use machine-translating fully automatic subtitles (from English to Chinese Simplified), with a low or a high level of non-verbal input. This study aims to provide some suggestions for machine translation improvements in automatic subtitling.

2. Methods and materials

2.1. Methods

The research combines objective and subjective measures to help further triangulate the experiment results and provide further insight into the subtitling production process. Three measures were used in this experiment, process logs, keystrokes, and a user experience questionnaire (explained in Figure 1).

For process logs, this experiment analysed the total time spent on post-editing through logs documents generated by a professional subtitle software[1], which is also used by the participants.

Windows Problem Steps Recorder (PSR)[2] is used to record keystrokes during the post-editing process. It is a tool provided by Windows to automatically capture steps on a computer. It is convenient for subtitlers because the experiment was conducted online. These recorded steps include insertions and deletions, which show subtitlers' revision behaviours. This study also considers the purpose of insertions and deletion since machine translation is only part of

[1] The software is achieved from https://www.1sj.tv/
[2] See https://learn.microsoft.com/en-us/office/troubleshoot/settings/how-to-use-problem-steps-recorder

automatic subtitling. It shows how much effort the subtitlers spend on machine translation in automatic subtitling. Therefore, WinMerge[3], detecting and displaying differences within text files, is used to compare the differences between post-editing automatic subtitles and original automatic subtitles.

The questionnaire in this study was adapted from the User Experience Questionnaire (UEQ) developed by Karakanta et al. (2022a) for end-user evaluation of MT in automatic subtitling.

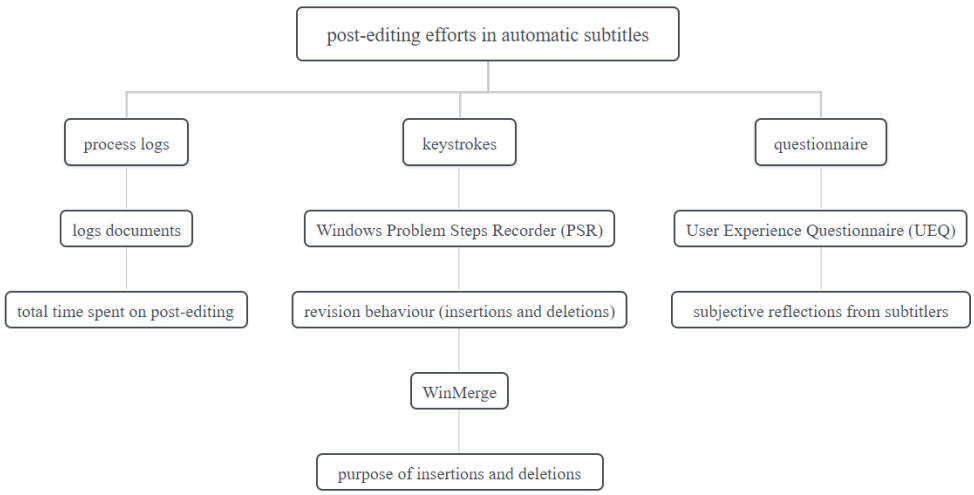

Figure 1. Three measures in this study.

2.2. Participants

12 subtitlers were recruited in the experiment. All are Chinese natives with English as their second language. All of them have passed the entry tests as a freelancer in a Mainland translation agency. According to the demographic data they provided, they have professional experience as subtitle translators in the relevant language pair, an average of 2.8 years (range 1-10 years). 75% of the participants have learned translation during their undergraduate or postgraduate study and 50% of them have passed the China Accreditation Test for Translators and Interpreters (CATTI) in the English-Chinese language pair. All of them have experience in using translation technologies such as machine translation. Over half of the subtitlers (58.3%) frequently use machine translation when they do translation projects, while just three participants seldom use it.

2.3. Materials

Video clips to be subtitled were selected based on the concept of "semiotic cohesion" (Díaz-Cintas & Remael, 2014, p. 51) and the research samples in Huang and Wang (2022). There are eight video clips in this experiment, four from a documentary film and four from TV series. All the video clips were cut from longer videos in English and each one lasts about one minute.

Table 1 shows the multimodal analysis of the image and speech information from the original materials to explain the selection.

[3] The software is achieved from https://winmerge.org/

	Description of the images	Description of the speech	Level of non-verbal input	Source material information
Documentary		Narration has no direct reference to image	Low (without subtitle-image and speech-gesture interaction)	<MINIMALISM: Official Netflix Documentary> by Netflix (2023)
Text 1	Moving shots of the minimalist going to work	Self-narration from a minimalist talking about his life		
Text 2	Moving shots of some houses	Narration from a third-person narrator talking about people's mistake of buying house		
Text 3	The screens when a minimalist faces the camera	Self-narration from a minimalist talking about his experience		
Text 4	Moving shots of a city	Narration from a third-person narrator talking about people's misunderstanding of buying		
TV series		Dialogues include pronouns that refer to the people in the images	High (with subtitle-image and speech-gesture interaction)	<Young Sheldon> Season 6 by CBS (2022)
Text 5	Static shots between six family members in the dining room	Diegetic dialogues between several people		Episode 2
Text 6	Static shots between four family members in their kitchen			Episode 3
Text 7	Static shots between six family members in the dining room			Episode 6
Text 8	Static shots between five family members in the dining room			Episode 12

Table 1. Multimodal analysis and selection criteria of the source materials.

	Total duration(s)	Number of sentences	Tokens
Text 1	72	18	115
Text 2	57	17	144
Text 3	65	15	116
Text 4	57	14	112
Text 5	60	37	174
Text 6	61	40	176
Text 7	62	33	189
Text 8	62	39	208

Table 2. Basic features of the source materials.

As shown in Table 1, Text 1-4 were selected from a documentary film, without subtitle-image and speech-gesture interaction. Therefore, they were evaluated as having a low level of non-verbal input because the images had no direct reference to the narration's content. Text 5-8, selected from a TV series, were considered to have a high level of non-verbal input. These videos contain character dialogues with facial expressions, pronouns, and gestures. The themes of all texts were based on the topic of life, especially daily life, to avoid any confounding results by different topics.

Each text has a similar duration and contains a complete scene to avoid any confusion. However, texts in the documentary and TV series were inevitably different in terms of their tokens and sentences (shown in Table 2).

All the subtitles were generated by the professional subtitle software through automatic transcription, spotting segmentation, and MT, without any human interruption. Besides, subtitlers received automatic subtitles in English for their references.

2.4. User experience questionnaire

An online questionnaire was used to collect subjective feedback from subtitlers. The objective of UEQ is to provide a user experience of post-editing and automatic subtitling. The questionnaire contained open and closed questions, which were delivered in English and were conducted through 问卷星 (www.wjx.cn). To obtain objective results, all responses were kept anonymous.

The questionnaire included three parts. The first part collected demographic data about the subtitlers, including gender, English proficiency, years of experience in subtitling and how often they use translation technologies, including machine translation. The second part focused on the user experience with the task of post-editing automatically generated subtitles. It contained 13 pairs of adjectives related to the post-editing experience for documentaries and TV series, in the form post-editing was... (difficult/easy, unpleasant/pleasant, etc.). Besides, it has evaluations on the quality of spotting and segmentation and the effort of editing them. For the second part, the author processed the scores using the formulae in the UEQ Data Analysis Tools (version 7)[4] to convert them to a scale of -3 to +3, with 0 representing a neutral mid-point. In the UEQ Data Analysis Tool, average scores between -0.8 and +0.8 are defined as neutral evaluations. Values below -0.8 correspond to negative and values above 0.8 to positive evaluations. The last part provided open questions on the quality and benefits of MT in automatic subtitling. Participants were also asked to provide their comments on machine translation and automatic subtitling.

[4] UEQ Data Analysis Tools: https://www.ueq-online.org/

2.5. Procedure

Before participating in the experiment, participants were asked to read the guidelines. The guidelines concern the task and the quality of subtitle production in the Code of Good Subtitling Practice (Ivarsson & Carroll, 1998). The quality guideline contains some major parts for subtitling, including grammar, spelling and punctuation, content and transfer, and readability. The guideline is accepted by researchers in AVT (Romero-Fresco & Pöchhacker, 2017; Huang & Wang, 2022), although it was released two decades ago. Besides, the participants were informed about the objective of the research, and the purposes of the data collection and gave their consent.

The subtitling tasks were carried out using the subtitling software and in one language pair: English to Chinese Simplified. Subtitlers had access to the internet as well as other resources normally used in their work. The participants were required to finish the tasks in one week. The experiment was conducted online.

The experiment includes two parts. In part one, the subtitlers were required to post-edit eight automatic subtitles. And they used the steps recorder when they were doing the post-editing tasks. In part two, the subtitlers were required to finish the user experience questionnaire and gave their comments on automatic subtitles and machine translation. All the subtitlers got their pay after finishing the task.

3. Data analysis

For each subtitler, the author collected the following data: 1) the final human post-edited subtitle files in SubRip.srt format; 2) logs documents from the subtitle tool, which records original and final timestamps; 3) keystrokes, using Windows Problem Steps Recorder (PSR), which automatically capture steps on a computer. At the end of the task, the subtitlers completed a questionnaire providing feedback on their user experience with automatic subtitling.

3.1. Process logs and keystroke logging

The time spent on post-editing was calculated based on the logging documents and steps recorder. Although each video clip lasts for about one minute, all clips have different tokens. Therefore, the average time spent on post-editing per token is calculated (see Figure 2). From Figure 2, subtitlers spent more time on just one text from TV series than texts from a documentary, while they spent less time on two texts from TV series than those from a documentary. Furthermore, in Text 1-4, the subtitlers spent about 8.75 seconds on post-editing per token. In Text 5-8, it takes about 8.60 seconds for the subtitlers to post-edit each token. Although Text 5-8 contain more non-verbal input, it seems that there are no significant differences (less than 0.5 seconds) between them and Text 1-4 in post-editing. This finding corroborates Huang and Wang's (2022) argument that non-verbal input from audiovisual texts was not an obstacle during post-editing.

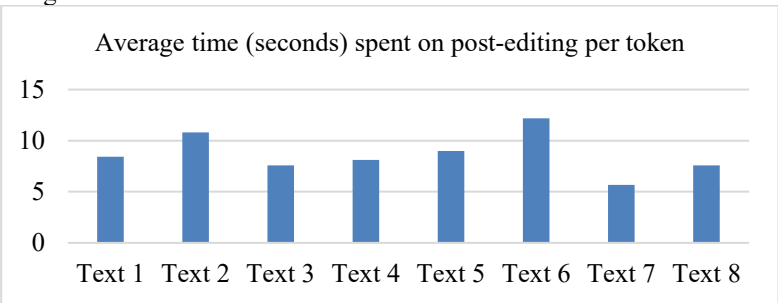

Figure 2. Average time spent on post-editing per token.

The number of keystrokes, insertions, and deletions were calculated by steps recorder (see Table 3). From Table 3, the deletions number of each text was much more than the number of insertions. Besides, when post-editing Text 5-8 from the TV series, the subtitlers used more insertions and deletions. Then, the subtitlers' revising behaviours were further analysed by WinMerge to get a full picture. WinMerge generated reports of differences between automatic subtitles and post-editing subtitles (samples are seen in Figure 3) and the reports showed the subtitlers revising efforts in translation, spotting, and segmentation. Considering that there are 12 subtitlers, the author calculated an average number of differences. Figure 4 shows the subtitlers' efforts on machine translation, spotting, and segmentation. It turns out that the subtitlers spent more effort on spotting and segmentation than on machine translation in Text 5-8. The reports also show how the subtitlers interact non-verbal information with verbal texts (samples are shown in Table 4). For instance, the subtitlers recognized characters through images and the audio in Text 6, so they revised the spotting, segmentation and translation. Without this interaction, the subtitles would make no sense.

	Average insertions	Average deletions
Text 1	44	111
Text 2	71	146
Text 3	45	109
Text 4	46	116
Text 5	76	205
Text 6	86	151
Text 7	69	135
Text 8	84	187

Table 3. Average insertions and deletions in different texts.

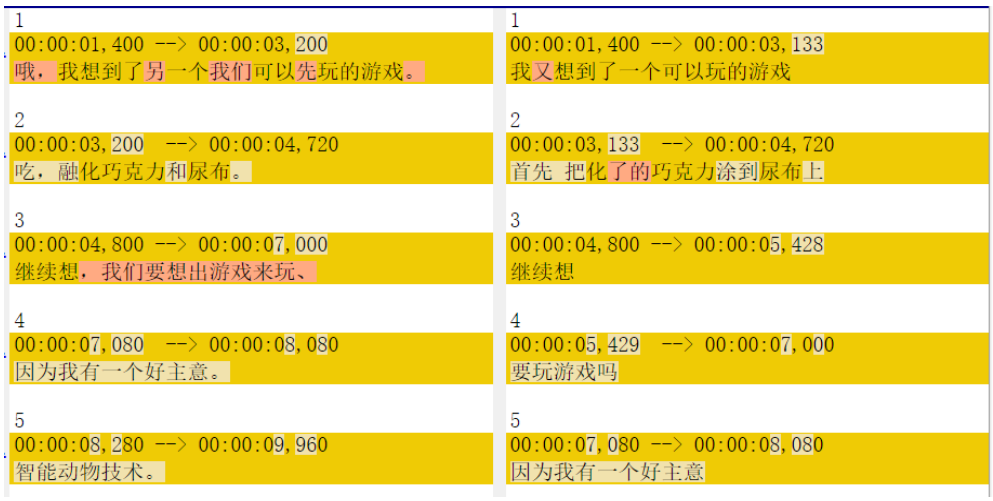

Figure 3. Screenshot of a WinMerge report from one subtitler in Text 8 post-editing.

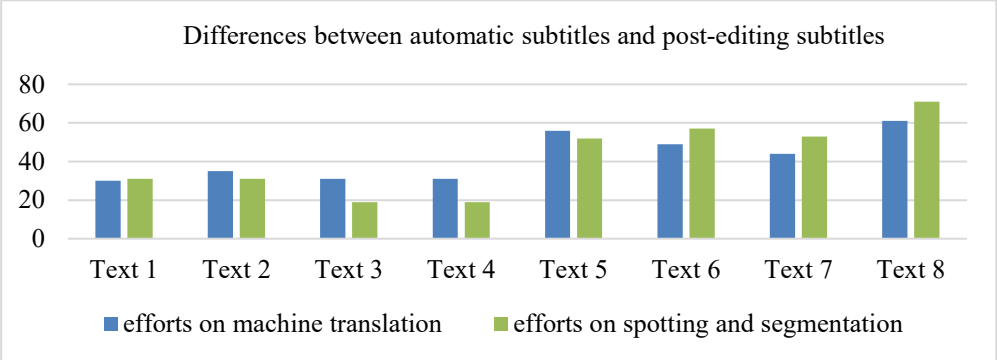

Figure 4. Differences between automatic subtitles and post-editing subtitles.

Text 6	Automatic subtitles	Post-editing subtitles
A multi-speaker event and pronouns in the text	13. 00:00:21,240 --> 00:00:**23,320**	13. 00:00:21,240--> 00:00:**22,333**
	只吃点吐司怎么样？**很好的吐司。**	只吃点吐司怎么样？
	14. 00:00:23,320 --> 00:00:23,560	14. 00:00:22,333 --> 00:00:23,560
	我可以的。	**好啊 吐司我会做**
	28. 00:00:42,070 --> 00:00:**43,630**	28. 00:00:42,070 --> 00:00:**43,333**
	我希望你能帮**我**解决**Sheldon的问题**。	我希望你能帮**我**解决
	29. 00:00:43,630 --> 00:00:44,470	29. 00:00:43,333 --> 00:00:44,470
	别管她了。	**谢尔顿** 别烦她了
	33. 00:00:49,190 --> 00:00:49,590	33. 00:00:49,190 --> 00:00:49,590
	我是什么？	**我**是什么来着？
	34. 00:00:49,790 --> 00:00:50,270	34. 00:00:49,790 --> 00:00:50,270
	没有盲文。	**侄辈母亲**

Table 4. Examples of the interaction between verbal and nonverbal information in a Win-Merge report.

3.2. Evaluation of user experience

The user experience (UX) scores are shown in Figure 5-6. Overall, the post-editing experience can be considered neutral to positive in Text 1-8 with different non-verbal input. The subtitlers found the post-editing process pleasant, enjoyable, and practical in all texts, with different levels of non-verbal input. When post-editing Text 5-8 with a high level of non-verbal input, the subtitlers found it more relaxed, exciting, fun, creative and motivating. When post-editing Text 1-4 with a low level of non-verbal input, the subtitlers felt less laborious, more efficient, simpler and faster, although there are no significant differences in average time spent on Text 1-4 and Text 5-8. In Figure 6, overall spotting and segmentation evaluations in all texts are neutral, except for the automatic segmentation evaluation. Compared with Text 1-4, the subtitlers considered automatic segmentation was poor in Text 5-8 from the TV series. The subtitlers found experience in spotting and segmentation much better when post-editing Text 1-4 with a low level of non-verbal input. These findings were in accord with the previous analysis of subtitlers' revising behaviours that they spent more effort in spotting and segmentation than translation.

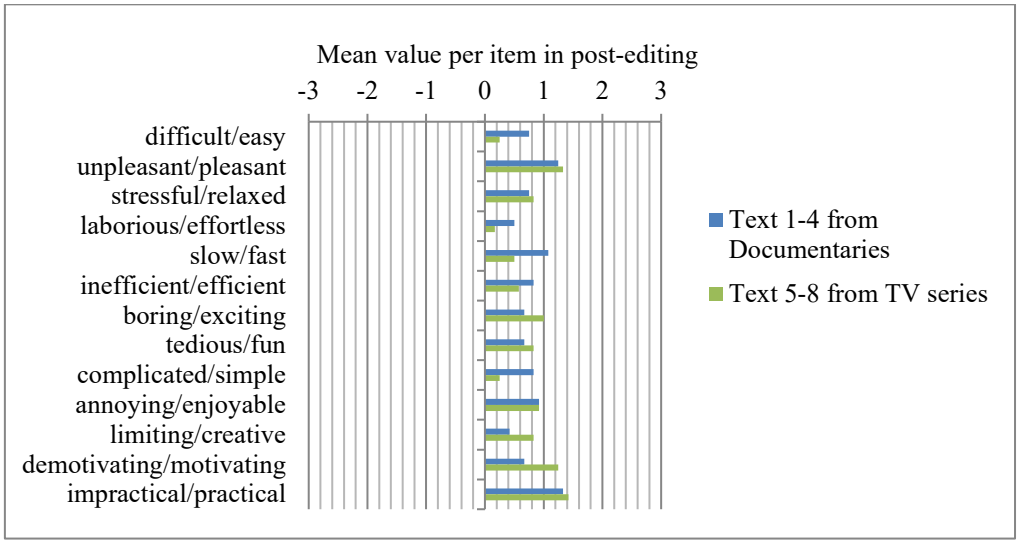

Figure 5. User experience (UX) scores in post-editing.

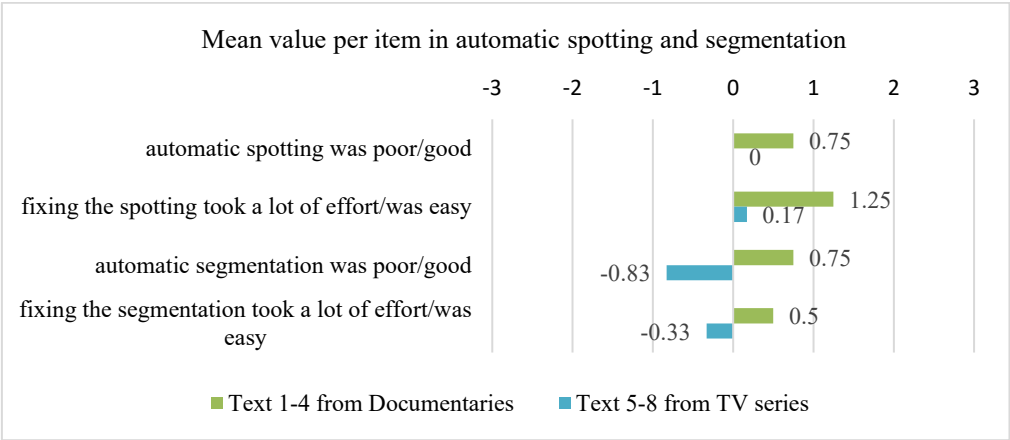

Figure 6. User experience (UX) scores in automatic spotting and segmentation.

3.3. Subtitlers' feedback

Main issues with the machine translation: For Text 1-4 from documentaries, some subtitlers found that for some sentences, the tense was wrong. Besides, some words were translated with the same meaning, although they occurred several times in one video. Two subtitlers thought that machine translation was not good at translating long sentences in subtitling. For Text 5-8 from the TV series, the subtitlers found more issues. For instance, the machine translation engine cannot recognize a multi-speaker event. The machine translations were literal and had problems with slang and new words created by the characters. At the same time, four subtitlers pointed out that the accuracy of machine translation was affected by the errors of automatic transcription. Most subtitlers responded that the translation style should be oral.

Main benefits of machine translation in subtitling: Most (67%) subtitlers mentioned that machine translation helps them to understand the main idea of the videos so that they can work efficiently. Some responded that machine translation helps them save time in typing. 91% of

the participants thought that machine translation helps the work of subtitlers while just one participant held the neural opinion.

Impression of using machine translation in subtitling: The subtitlers gave feedback on the danger for the profession of the subtitler from using machine translation. Four participants said that there is no danger for the subtitlers. Half of the participants mentioned that they may rely on machine translation if they use it more frequently. They may lack initiative and think less when they become accustomed to using machine translation. Two participants predicted that the subtitlers who work with general texts may be replaced in the future.

4. Conclusions

This study examined post-editing efforts in automatic subtitling, with a focus on non-verbal input's effect on machine translation. In general, time spend is not significantly different between videos with low and high levels of non-verbal input, although subtitlers felt less laborious when translating texts with less non-verbal information. Furthermore, the subtitlers spent more effort on revising spotting and segmentation than translation when they post-edited texts with more non-verbal information. It may help to explain why the subtitlers felt faster when they translated texts with a low level of non-verbal input. The comparison between automatic subtitles and post-editing subtitles also shows that the subtitlers revised translation, spotting and segmentation to interact non-verbal information (images and audio) with verbal information (texts). Machine translation had more problems with texts containing non-verbal input, according to subtitler feedback. Most subtitlers hold a positive attitude towards machine translation usage. However, subtitlers may rely on it if they use it more frequently.

This experiment also offers valuable insights for MT improvements in automatic subtitling. For instance, pronoun detection in TV series and more high-quality training data with non-verbal information are needed to improve the machine translation engine. Additionally, it is crucial to provide more training to students or subtitlers to reduce reliance on machine translation during subtitle creation.

However, the study has certain limitations. The experiment was not conducted on a large scale, involving only 12 subtitlers, and the video clips were limited to one documentary and TV series. Further research, through eye tracking and interview, is necessary to explore how subtitlers interact non-verbal information with verbal text to help them post-edit machine translation in automatic subtitling.

Acknowledgement

The author kindly thanks all the subtitlers who took part in the experiment.

References

Bellés-Calvera, L., & Quintana, R. C. (2021). Audiovisual translation through NMT and subtitling in the Netflix series 'cable girls'. In R Mitkov et al. (Eds). *Proceedings of the Translation and Interpreting Technology Online Conference* (pp. 142-148). INCOMA Ltd.

Díaz-Cintas, J., & Remael, A. (2014). *Audiovisual translation: Subtitling*. Routledge.

Georgakopoulou, Y. (2021, March 22). Implementing Machine Translation in Subtitling. Multilingual. https://multilingual.com/implementing-machine-translation-in-subtitling/

Guillot, M.-N. (2018). Subtitling on the cusp of its futures. In L. Pérez-González (Ed.), *The Routledge handbook of audiovisual translation* (pp. 31–47). Routledge.

Huang, J., & Wang, J. (2022). Post-editing machine translated subtitles: examining the effects of non-verbal input on student translators' effort. *Perspectives, Studies in Translatology, ahead-of-print(ahead-of-print)*, 1–21.

Ivarsson, J., & Carroll, M. (1998). *Code of good subtitling practice.* European Association for Studies in Screen Translation.

Karakanta, A. (2022). Experimental research in automatic subtitling: At the crossroads between machine translation and audiovisual translation. *Translation Spaces, 11*(1), 89–112.

Karakanta, A., Bentivogli, L., Cettolo, M., Negri, M., & Turchi, M. (2022b). Towards a methodology for evaluating automatic subtitling. In H. Moniz et al. (Eds). *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 333-334). European Association for Machine Translation.

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

Koponen, M., Sulubacak, U., Vitikainen, K., & Tiedemann, J. (2020a). MT for Subtitling: MT for Subtitling: Investigating professional translators' user experience and feedback. In J. E. Ortega, M. Federico, C. Orasan, & M. Popovic (Eds), *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation* (pp. 79-92). Association for Machine Translation in the Americas.

Koponen, M., Sulubacak, U., Vitikainen, K., & Tiedemann, J. (2020b). MT for subtitling: User evaluation of post-editing productivity. In A. Martins et al. (Eds). *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 115–124). European Association for Machine Translation.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358-392.

Pérez-González, L. (2014). Multimodality in translation and interpreting studies: Theoretical and methodological perspectives. In S. Bermann, & C. Porter (Eds.), *A companion to translation studie*s (pp. 119–131). Wiley Blackwell.

Perego, E. (2009). The codification of nonverbal information in subtitled texts. *In New trends in audiovisual translation* (pp. 58–69). Cromwell.

Romero-Fresco, P., & Pöchhacker, F. (2017). Quality assessment in interlingual live subtitling: The NTR model. *Linguistica Antverpiensia, New Series: Themes in Translation Studies, 16*, 149–167.

Saarikoski, L., Van Rijsselbergen, D., Hirvonen, M., Koponen, M., Sulubacak, U., & Vitikainen, K. (2020). MEMAD Project: End User Feedback on AI in the Media Production Workflows. *Proceedings of IBC 2020*.

Varga, C. (2021). Online Automatic Subtitling Platforms and Machine Translation. An Analysis of Quality in AVT. *The Scientific Bulletin of the Politehnica University of Timişoara. Transactions on Modern Languages. 20* (1), 37-49.