Post-editing in Automatic Subtitling: A Subtitlers' Perspective

Alina Karakanta[1,2], Luisa Bentivogli[1], Mauro Cettolo[1],
Matteo Negri[1], Marco Turchi[1]
[1]Fondazione Bruno Kessler [2]University of Trento
{akarakanta,bentivo,cettolo,negri,turchi}@fbk.eu

Abstract

Recent developments in machine translation and speech translation are opening up opportunities for computer-assisted translation tools with extended automation functions. Subtitling tools are recently being adapted for post-editing by providing automatically generated subtitles, and featuring not only machine translation, but also automatic segmentation and synchronisation. But what do professional subtitlers think of post-editing automatically generated subtitles? In this work, we conduct a survey to collect subtitlers' impressions and feedback on the use of automatic subtitling in their workflows. Our findings show that, despite current limitations stemming mainly from speech processing errors, automatic subtitling is seen rather positively and has potential for the future.

1 Introduction

Machine Translation (MT) is today widely adopted in most areas of translation and post-editing has been established as a professional practice, shaping the landscape of the translation industry. Audiovisual Translation (AVT) is one area where MT has for long found limited success (Burchardt et al., 2016). Among the main reasons are the inability of MT systems to deal with creative texts (Guerberof-Arenas and Toral, 2022) and the multimodality of the source, since the translation depends on visual, acoustic and textual elements (Taylor, 2016). For subtitling, additional challenges are posed by the formal requirements of the target: subtitles should not exceed a specific length and should be synchronised with the speech (Carroll and Ivarsson, 1998). However, recent developments in neural machine translation (NMT) and speech translation (ST) are paving the way for viable and usable (semi-)automatic solutions for subtitling. Compared to solutions providing MT for subtitling, automatic subtitling tools do not simply translate human-generated source language subtitles, but incorporate automatic transcription of the speech, MT, automatic synchronisation (spotting) and segmentation of the translated speech into subtitles. Altogether, these technologies come with the promise of reducing the human effort in the subtitling process, but, to date, automatic subtitiling has still to be put to test by the actual users.

Even though translators are fundamental for the advance of new technologies, their views are often not sufficiently considered (Guerberof-Arenas, 2013). The study of subtitlers' perceptions of the technology they are interacting with can be beneficial for all stakeholders in the AVT industry. Furthermore, the inclusion of subtitlers in the process of technological change can alleviate their resistance to adopting technologies (Cadwell et al., 2018). Developers can direct their implementation efforts in the right direction to provide user-friendly tools and interfaces (Moorkens and O'Brien, 2017), and AVT trainers can identify necessary skills for teaching and training (Bolaños-García-Escribano et al., 2021). A better understanding of subtitlers' interaction with technology can help define the rising profession of the subtitler post-editor (Bywood et al., 2017), and establish metrics and standards to protect subtitlers against dropping rates and ensure fairness (Nikolic and Bywood, 2021).

In response to the challenges brought about by increasing technologisation, in this work we conduct a survey of subtitlers' perspectives on the developing paradigm of automatic subtitling. This survey is a timely contribution to take stock of this nascent technology and its implementation in the subtitling profession from the very beginning, while setting the stage for further developments. The survey focuses on the subtitlers' user experience when post-editing automatically-generated subtitles from and into different Western European languages. It also aims at collecting feedback on the main issues and benefits of the technology, as well as on the impact of automatic subtitling on the subtitler's profession. Based on qualitative and quantitative analysis of a survey questionnaire, we provide a participant-based evaluation of automatic subtitling and a comprehensive view of subtitlers' attitudes towards this new paradigm. Our findings indicate that despite its current limitations mainly related to challenges in speech processing, automatic subtitling has potential and its benefits are already recognised by the users. Based on the received criticisms, we provide a list of recommendations for future improvements in automatic subtitling tools, which we hope will serve as a guide for technology developers. We further release the questionnaire and responses to foster replication and reproducibility in automatic subtitling.[1]

2 Related work

Automatising subtitling has recently received growing interest. One research direction aims at controlling the generation of captions and subtitles based on particular variables and properties, such as genre (Buet and Yvon, 2021), length (Lakew et al., 2019; Liu et al., 2020) or alignment between source and machine-translated subtitles (Cherry et al., 2021). Though relevant from the technological standpoint, this line of research has employed automatic metrics for the evaluation of MT and has not included subtitlers in the evaluation process.

Other studies have tested the usability of MT for subtitles by focusing on quality and productivity, mainly through the task of post-editing (PE). The human evaluation, however, did not always involve professional subtitlers. Some studies used volunteers (C. M. de Sousa et al., 2011), native speakers (Popowich et al., 2000; O'Hagan, 2003) or translators (Melero et al., 2006). Nevertheless, subtitling requires special training and skills which native speakers or translators do not necessarily possess. Larger scale evaluations involved professional subtitlers, but focused on machine translating human-generated source language subtitles. This setting has less challenges than automatic subtitling, since the source text is error-free and already compressed, while the spotting and segmentation are performed by a human. Volk et al. (2010) built an MT system between Scandinavian languages, which was tested by professional subtitlers, and collected their feedback in a non-structured way. The large-scale SUMAT project (Etchegoyhen et al., 2014) involved two professional subtitlers per language pair, who performed post-editing and rated their perceived PE effort. Matusov et al. (2019) evaluated the productivity gains of their proposed English into Spanish system with two post-editors, who were additionally asked to rank the adequacy, fluency and design of the subtitles. User feedback was collected in a non-structured way, where subtitlers commented on the post-editing process and on their perception of MT in their workflows. Lastly, Koponen et al. (2020b) performed a comprehensive human evaluation of their MT systems for Scandinavian languages. The evaluation included the collection of product and process (keystrokes) data, as well as rich feedback based on a mixed methods approach using questionnaires and semi-structured interviews.

Our present study builds upon the work by Koponen et al. (2020b) by extending the feedback collection to a larger participant sample (22 compared to 12) working in a variety of Western European language pairs. One main difference is the technology behind the generation of the target subtitles. In our study, respondents are asked to evaluate their user experience after post-editing subtitles generated through a three-step fully automatic process involving transcription, synchronisation and translation. On the contrary, in (Koponen et al., 2020b) source subtitles were first obtained by a human (subtitle template), and then machine translated and aligned to the original frames. In addition, the subtitlers used their preferred subtitling software in the PE tasks. However, as the authors admit, the subtitling tools are not designed for MT Post-editing (MTPE), and may therefore not be optimal for the task. Our work has the benefit of evaluating the PE experience using a professional tool specifically tailored for post-editing automatically generated subtitles as a case study.

[1] https://github.com/fatalinha/subtitlers-have-a-say

3 Methodology

The survey described in this paper was conducted in December 2021 and consisted in respondents filling in a questionnaire after having taken part in testing sessions of an automatic subtitling tool.

3.1 The task

In the PE task, subtitlers were required to post-edit the automatically-generated subtitles of 8 video clips. The clips were self-contained excerpts from different TV series (drama), each around 3 minutes long, amounting to a total duration of 30 minutes. TV series were selected as the material to post-edit since they are representative examples of real subtitling tasks. In addition, they contain elements which are particularly challenging both for human subtitlers and automatic systems, such as background noise, slang, overlapping speech and multi-speaker events. The original language of the series was English. Since all subtitlers edited the same clips but not all of them worked with English as source language, we used the dubbed version for subtitlers working from Spanish and Italian.

The task was performed over two consecutive days and the subtitlers took sufficient breaks between each video to avoid fatigue effects. The subtitlers worked from their personal office without any explicit time limit. Before starting the task, all participants, regardless of their previous experience with the subtitling tool, were asked to familiarise themselves with it by watching a video tutorial, in which the functionalities of the tool were explained. This setting resulted in a homogeneous task for all participants, with a sufficient duration to develop reliable judgements and a robust opinion on their user experience.

3.2 The tool

The automatic subtitling system selected for this study is integrated in a novel subtitling tool, Matesub.[2] Matesub is a typical instance of an automatic subtitling tool. It features a state-of-the-art ST system, with automatic generation of timestamps for the translated subtitles – a process called automatic spotting (or auto-spotting) – and automatic segmentation of the translated audio into subtitles.

Figure 1 shows a screenshot of the tool. The subtitlers are presented with a list of the automatically generated subtitles (upper left box) and the video on which the subtitles appear (upper right). The boxes corresponding to each subtitle appear at the bottom of the screen, superimposed on a waveform which allows the subtitler to identify parts of the video corresponding to the selected speech segments. The position and length (duration) of the boxes can be adjusted to match the beginning and the end of the spoken utterance and to accommodate the time the subtitle will appear on screen. Moreover, the tool has a quality assurance feature which raises an issue whenever pre-defined subtitling constraints are violated, for example if a subtitle is too long (length) or disappears too early (reading speed). All these elements, along with other useful features, such as keyboard shortcuts and positioning or colour settings, are implemented in most subtitling editors not offering MT integration, therefore post-editing subtitles in Matesub has the benefit of being representative of subtitlers' real working settings. The tool is free, tested in real-life use cases and is already being used by professional subtitlers.

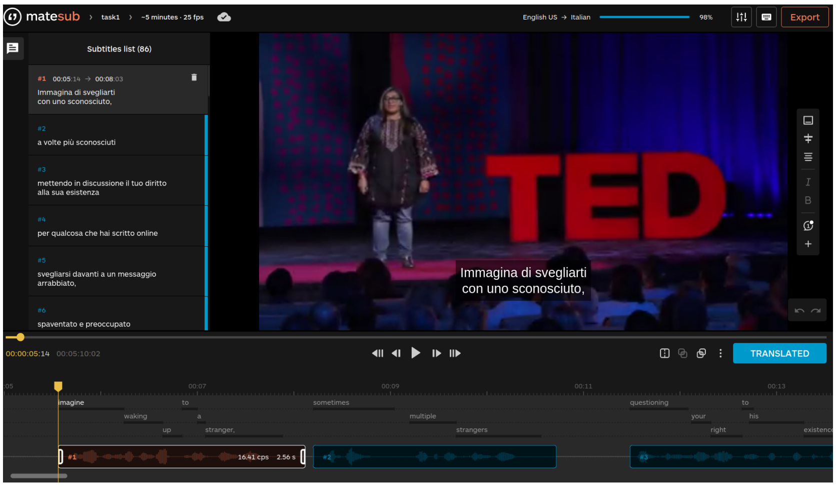
3.3 Respondents

The respondents were professional subtitlers who took part in the post-editing task with the Matesub tool. They were recruited through a language service provider (Translated.com). Participation to the survey was voluntary and the responses were collected anonymously. Before starting the survey, participants were informed about the objective of the research, the purposes of the data collection and gave their consent. In total, 22 out of 24 subtitlers responded to the questionnaire (91% response rate). The subtitlers worked in different language pairs. Table 1 shows the number of subtitlers for each language pair. Subtitlers worked in from-English, into-English, but also non-English language pairs, which are often disregarded in MT research (Fan et al., 2021). The focus of the survey is to obtain a broad overview of subtitlers' opinions on automatic subtitling, regardless of the language-specific performance of the technology. Therefore we opted for selecting respondents so as to cover a wide range of language pairs.

3.4 Survey and questionnaire

The questionnaire was set up as an online form containing open and closed questions. It was delivered in English for all respondents and contained three parts. The first part collected factual information about the subtitlers, such as years of experience in subtitling, years of experience in MTPE

[2] `https://matesub.com/`



#1 00:05:14 → 00:08:03
Immagina di svegliarti
con uno sconosciuto,

#2
a volte più sconosciuti

#3
mettendo in discussione il tuo diritto
alla sua esistenza

#4
per qualcosa che hai scritto online

#5
svegliarsi davanti a un messaggio
arrabbiato,

#6
spaventato e preoccupato

Immagina di svegliarti
con uno sconosciuto,

Figure 1: The Matesub subtitling tool.

Language pair	Subtitlers
Spanish → English	2
Spanish → Italian	3
Spanish → German	3
Italian → French	3
English → French	2
English → Spanish	3
English → Polish	3
English → Dutch	3

Table 1: Respondents per language pair.

and how often they use Matesub. Three questions focused on the working settings and the diffusion of MT in subtitling jobs. These questions asked how often their subtitling jobs involved using master templates, working directly from the video, and editing machine translated subtitles.

The second part of the questionnaire focused on the respondents' user experience with the task of PE automatically generated subtitles. We used the User Experience Questionnaire (UEQ) by Koponen et al. (2020a), a version of the UEQ of Laugwitz et al. (2008) for end-user evaluation of software products, which has been adapted to post-editing experience. This selection of questionnaire facilitates comparison of PE in automatic subtitling with the PE experience based on a different system. By using an existing questionnaire, we respond to the need for standardisation in experimental research in AVT and MT. The questionnaire contained 13 pairs of adjectives related to the post editing experience, in the form *Post-editing was... (difficult/easy, unpleasant/pleasant, stressful/relaxed, labourious/effortless, slow/fast, inefficient/efficient, boring/exciting, tedious/fun, complicated/simple, annoying/enjoyable, limiting/creative, demotivating/motivating, impractical/practical*). Since the tool features auto-spotting and automatic segmentation, we included evaluations on the quality of spotting and segmentation and the perceived effort of editing them. The responses are provided on a scale of -3 to +3, with 0 representing a neutral mid-point. As in the UEQ, average scores between -0.8 and +0.8 are considered neutral evaluations, while scores below -0.8 correspond to negative evaluations and scores above 0.8 to positive evaluations.

The last part of the questionnaire contained open questions on the quality of MT, auto-spotting and automatic segmentation, as well as the subtitlers' opinion on the benefits of automatic subtitling, whether it helps the work of subtitlers and whether they see any dangers for the profession of subtitlers from using automatic subtitling. The open questions were analysed based on thematic analysis (Braun and Clarke, 2006) using the Taguette[3] software. This analysis aimed at identifying main issues with the technologies implemented in the tool, as well as the main benefits from using automatic subtitling. The general opinion on usability is coded as positive, neutral/mixed or negative.

4 Results

4.1 Subtitlers' profiles and working settings

The respondents had on average 2.3 years of experience as subtitlers (SD=1.5, range 1-5 years) and 2.6 years of experience with MTPE (SD=2.4, range 0-10 years). In terms of working settings, there is large variability in the way subtitling is per-

[3] `https://www.taguette.org/`

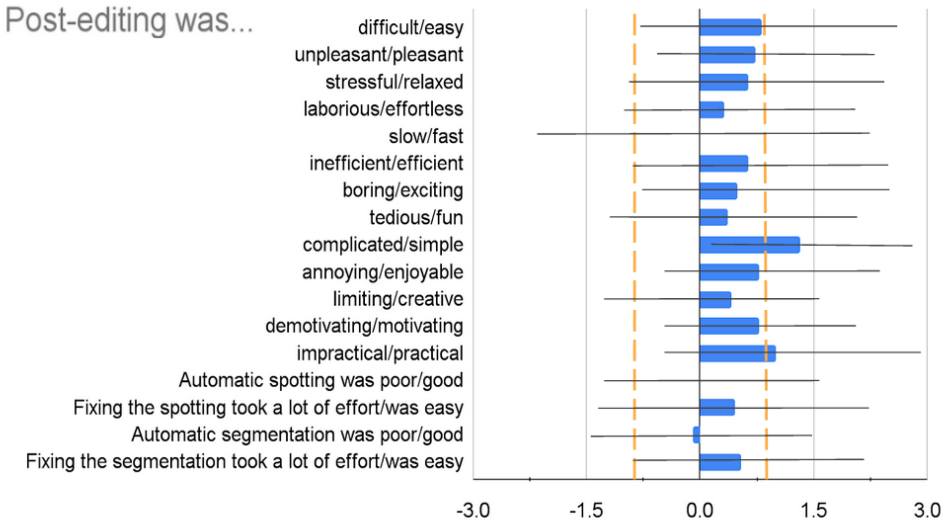


Figure 2: User experience (UX) scores. Interrupted vertical lines mark the -0.8/+0.8 threshold for neutral evaluations. Horizontal lines mark standard deviation.

formed. To the question *How often do your subtitling jobs involve master templates*, 5 subtitlers responded they never work with templates, 4 rarely, 6 sometimes and 7 often. When asked *How often do your subtitling jobs involve working directly from the video*, 3 subtitlers responded that they always work from the video, 4 often, 6 sometimes, 5 rarely and 4 never. When it comes to the question *How often do your jobs involve editing machine-translated subtitles*, 4 subtitlers mentioned that they always edit machine-translated subtitles, 3 often, 4 sometimes, 6 rarely and 5 never. This shows that there is variability in the professional conditions in subtitling when it comes to the use of tools, settings and requirements but, despite this, MT is a reality for subtitling. In addition, the responses confirm that our respondent sample covers different levels of expertise and a broad skill range.

4.2 User experience

The mean scores for the user experience across subtitlers and language pairs are shown in Figure 2. Overall, the post-editing experience can be considered as neutral to positive, with all except one mean scores leaning on the positive side of the scale. The subtitlers found the post-editing process simple and practical. Even though still in the neutral range, the lowest scores were observed for the quality of autospotting and automatic segmentation, where mean scores are close to 0.

When comparing the scores with the study of Koponen et al. (2020a), our scores are more distributed towards the positive side, even though a direct comparison of the user experience of the different subtitling systems is not the focus of this paper. It should also be noted that our sample is larger (22 respondents instead of 12) and with a larger variety in language pairs (8 compared to 4). In (Koponen et al., 2020a), the lowest average scores were found for the adjectives *laborious/effortless* and *limiting/creative*. This adjective pairs received low scores in our study too, however with *slow/fast* having the lowest score and a very large deviation. Similarly, the quality of autospotting and segmentation had lower scores than the effort to fix them. All in all, the user experience scores show that PE in automatic subtitling is a task found acceptable by the subtitlers and pointed out particular limitations, mainly related to the technical aspects of spotting and segmentation.

4.3 Subtitlers' feedback

Main issues with automatic subtitling Table 2 shows the main issues for automatic translation, auto-spotting and segmentation, as identified based on the thematic analysis of the subtitlers' responses to the open questions. For automatic translation, speech recognition errors seem to be the most common reason for errors in the translation (10 statements). Subtitlers mentioned that translation quality was highly influenced by the speaker's accent, audio quality and the speed of speech. For example, they mentioned that *muffled or fast speech, music and background noises can often confuse the AI*. Speech recognition errors have indeed been identified as the main issue for speech

Automatic translation		Autospotting		Segmentation	
Speech/audio recognition errors	10	Inaccurate (starting to early, too late)	10	Oversegmentation (too may short subtitles)	6
Lexical, punctuation, case	7	False negatives (no subtitle when speech)	5	No respect of syntactic/semantic units	5
Missing context, inconsistencies	5	False positives (subtitle when no speech)	3	No respect of constraints and guidelines	4
		Not respecting visual elements (shot changes)	2	Undersegmentation (too long subtitles)	3
Worked well	3	Worked well	6	Worked well	5

Table 2: Main issues related to automatic translation, autospotting and segmentation, and number of statements.

translation systems, regardless of whether they are direct or cascaded architectures (Bentivogli et al., 2021). The second group contained lexical errors, such as the translation of slang, idioms, colloquial expressions, figurative language and named entities, and in some cases, casing and punctuation (7 statements), with subtitlers reporting that *automatic translation still tends to be a bit too literal*. Translations out of context or words translated individually or inconsistently across the video were also mentioned as common issues (5 statements). A subtitler noted that inconsistent translation suggestions by the system *may lead the human translator to lose consistency as well*. Three subtitlers thought translation worked well.

For autospotting, lack of accuracy was the main reported issue (10 statements), since subtitlers thought that subtitles often started too early or too late and were not properly synchronised with the speaker. False negatives (no subtitle created when there is speech) and false positives (subtitles created when there is no speech) were also reported in 5 and 3 statements respectively. All these factors are related to common speech recognition issues, for example when speech is not recognised due to bad audio quality or when background noise is recognised as speech. Some subtitlers (2 statements) mentioned that automatically-spotted subtitles did not respect shot changes and other visual elements. Six subtitlers reported that autospotting worked pretty well or did not report any issues.

For automatic segmentation, oversegmentation (unnecessarily segmenting subtitles into small pieces) and undersegmentation (failing to segment too long subtitles) were mentioned in 6 and 3 statements respectively. Other issues were that the segmentation did not respect the norms of the target language because of splitting semantic/syntactic units (5 statements), and that segmentation resulted in subtitles not respecting the guidelines and length/reading speed constraints.[4] Five subtitlers affirmed that automatic segmentation worked well.

Main benefits of automatic subtitling When asked about the main benefits of automatic subtitling, *speed* was considered the main benefit by almost all subtitlers (18/22). Surprisingly, this is in contrast with the low mean score for slow/fast in the UX questionnaire. When looking into the benefits reported by subtitlers who rated the PE experience as slow (negative values for *slow/fast*), all of them mentioned that it saves time, but only on the creation of subtitle boxes and setting the timestamps. This shows the importance of not relying only on quantitative scores in participant-based studies, but complementing the judgements with quantitative explanations. Additionally, *efficiency* was noted as a benefit in 10 statements and *reduction of effort* related to technical aspects in 6 statements. Specifically, subtitlers reported that automatic subtitling *saves a lot of tedious work*, *creates a guideline of what needs to be translated instead of watching the whole video* and *serves as a starting template*, which, as a result, *allows focusing more on the translation* rather than having to spend time on technical aspects. The provision of *useful suggestions* was mentioned in 2 statements, related to subtitling solutions that the subtitler had not considered or to terminology and vocabulary.

General impressions for the subtitling profession To the question whether they think that automatic subtitling helps the work of subtitlers, 14 subtitlers responded positively, 5 gave neutral/mixed statements and 3 claimed that in most cases automatic subtitling does not help. The subtitlers who responded neutrally mentioned as concerns that the quality depends on the language,

[4] Netflix guidelines: https://partnerhelp.netflixstudios.com/hc/en-us/articles/360051554394-Timed-Text-Style-Guide-Subtitle-Timing-Guidelines

audio quality, and that it may be useful only for some applications (e.g. *template* creation, *other audiovisual products, such as online conferences or courses, documentaries*).

When asked whether they see any possible danger to the profession because of automatic subtitling, 8 subtitlers mentioned they see no dangers at all, 8 subtitlers saw no dangers for the time being, given the current state of the technology and its low diffusion, while 9 subtitlers identified some type of danger. Possible dangers were the loss in the quality of the final subtitles (4), dropping rates (2) and having less or no work if clients select cheaper, automatic options (5). Another danger identified was the improper application of the technology (3 statements), where subtitlers considered that the profession is not at risk only as long as a human is involved in the final phase.

5 Discussion

This study focused on subtitlers' user experience and perspectives on the task of post-editing automatically generated subtitles. Our findings suggest a neutral to positive experience. Even though there are those who still see no benefits from this new technology, automatic subtitling was welcomed with enthusiasm by many subtitlers, as an aid to save time and effort. As with studies on MTPE experience (Guerberof-Arenas, 2013; Bundgaard, 2017), subtitlers have expressed disfavour towards automatic subtitling in respect to technological flaws, but also acknowledged its positive aspects and expected technology to shape their profession in the near future. As for the dangers to the profession, most criticisms were not rooted in the fear of being outperformed by automatic systems, but rather in the effect of technology on the final product and market consequences (Vieira, 2020). The positive aspects of technology can only be appreciated when combined with respectful and ethical professional and market practices.

Previous work reporting feedback of subtitlers focused on a setting where MT was applied to human-generated subtitles. The views of the subtitlers involved did not lead to auspicious conclusions in favour of the use of MT in subtitling. In spite of encouraging automatic evaluation scores, subtitlers were cautious in reporting productivity gains in (Volk et al., 2010), while in (Etchegoyhen et al., 2014) PE experience was rated as rather negative (2.37 on a 1-5 scale), with MT being useful only for simple and short sentences. An increase in productivity for simple sentences was reported in (Matusov et al., 2019), where the two subtitlers rated their experience as fair. In (Koponen et al., 2020a) the participants did not find PE particularly difficult but characterised it as negative or limiting and did not think MTPE increased productivity. Similar criticisms were reported for MT quality in our study, with MT described as too literal, unable to properly translate spoken and figurative language. However, most subtitlers acknowledged that automatic subtitling makes their work faster and more efficient, especially when compared to *old-style subtitling*. The difference of our study compared to studies of MT for subtitling is the automatisation not only of the translation, but also of the technical aspects of spotting and segmentation. Subtitlers recognised the importance of automatising these aspects, which are often characterised as tiresome and dull. By not focusing only on the translation but the automation of the technical aspects, automatic subtitling allows subtitlers to spare time and effort on the tedious part of the work (spotting and segmentation) and unleash their creativity in adjusting the final text.

Our study aimed at providing a broad view of subtitlers' perspectives, by complementing quantitative scores with open questions, attempting to cover several language pairs and a range of subtitler profiles. However, we acknowledge that the findings should be interpreted with some caution. Questionnaire-based studies have a context-bound nature and may be affected by factors such as the system (quality, language), the participants (age, familiarity with technology) and the setting (Tuominen, 2018). Therefore, some limitations should be considered when drawing conclusions.

Firstly, responses and user experience scores may have been affected by the language pair, due to differences in the subtitling quality depending on the ASR and MT performance, despite keeping all other settings (videos, instructions) equal. Still, we opted for not reporting results separately for each language pair, since the sample size per pair (2-3) would be too small to draw robust and generalizable conclusions on a per-language basis. Second, even though we attempted to include a broad range of professional subtitler profiles, the group is not necessarily representative of the subtitlers' general population. For example, the respondents' age, a variable not collected in our survey, may affect their technological acceptance. Moreover, their experience in subtitling, template translation and MTPE varies. We found in statistical tests that the only variable affecting the user experience is MTPE experience. Subtitlers with less experience ($<=$ 2 years) had significantly higher user experience scores than the more experienced ones.[5] It is possible that experts, already being used to a certain level of MT output quality and to their preferred interfaces, are less willing to change tasks and tools, while novices, having less consolidated working practices, are more open and less critical against new interfaces and workflows. Accepting to take part in a task involving automatic subtitling already means the subtitlers were willing, curious or even familiar with the technology, and therefore may have been positively inclined towards automatisation in subtitling, contrary to many AVT professionals (Audiovisual Translators Europe, 2021).

Lastly, the interface used in PE has a great influence on user experience. We selected Matesub as a typical instance of an automatic subtitling tool. However, the generalisability to other tools is not guaranteed. In an attempt to test whether previous experience with Matesub had an effect on user experience, we separated the respondents in two groups based on their responses to the question *How often do you use Matesub in your subtitling jobs*: regular users (often, sometimes) and occasional (never, rarely). We found that familiarity with the tool did not have an effect on the average user experience scores.[6] This shows that the tool is user-friendly, with a steep learning curve, and does not require extensive training. Less user-friendly tools may negatively affect the post-editing experience. Despite these limitations, this study presents a screenshot of the current state of the quickly evolving technology, necessary to drive implementation efforts in the right direction.

5.1 Recommendations for improvement

Our findings have identified some limitations of current automatic subtitling systems. Based on the subtitlers' feedback, we present a list of suggestions for improving automatic subtitling tools in a direction that benefits the user experience. The suggestions are listed in order of priority.

• **Improving autospotting and segmentation**. The main benefit of automatic subtitling according to the subtitlers was eliminating tedious work and leaving more space for creativity. Given that many criticisms were addressed to the quality of autospotting and segmentation, improvements in the automation of technical aspects are a priority. Except for improving the accuracy of autospotting through enhanced audio processing and a more syntactically-informed segmentation, interaction with these elements could become more user-friendly. For example, it could be useful to implement interactive features such as automatic adjustment of subtitle boxes to match length and reading speed constraints after subtitlers translate or finish editing one subtitle.

• **Improved audio pre-processing**. Most problems in the translation, autospotting and segmentation stemmed from the segmentation of the audio. This is an open problem in speech processing (Gaido et al., 2021; Tsiamas et al., 2022); audio segmentation is typically approached by breaking the audio on speaker silences, considered as a proxy of clause boundaries, and not on syntactic information. A syntax-unaware segmentation is responsible for translations out of context and the issues in segmentation (over-undersegmentation, no respect of syntactic units). In addition, the reported cases of false positives/false negatives in autospotting (see Table 2) indicate that voice activity detection technologies should be improved to properly distinguish speech from noise.

• **Improving in-video consistency**. Consistency of MT suggestions is important for easily spotting errors and for avoiding repetitive corrections. Consistency can be improved through adaptive MT (Biçici and Yuret, 2011) or document-level MT (Lopes et al., 2020).[7] Another direction could be the integration of external resources, such as termbases and translation memories. These aids have passed the test of time and are usually the first requirement of users before overshooting with MT solutions (Audiovisual Translators Europe, 2021).

• **User experience vs. automatic metrics**. Punctuation and casing was reported as an issue for automatic translation. However, WER, the metric used to evaluate ASR systems, is normally computed in a case/punctuation insensitive way. Casing and punctuation cannot be derived directly

[5] Novices (N=14, M=1.0, SD=0.7) vs Experts (N=8, M=-0.4, SD=1.1). Based on an equal-variance independent samples t-test: ($t(20) = 3.82, p = .001$)

[6] Regular (N=14, M=0.6, SD=1.2) vs Occasional (N=8, M=0.4, SD=0.9). ($t(20) = 0.42, p = .679$)

[7] However, it should be noted that (Koponen et al., 2020b) found no preference for document-level MT compared to sentence-level MT in subtitling.

from the audio and therefore these errors are traditionally considered as less relevant by the scientific community. On the contrary, in the context of automatic subtitling they must be weighed appropriately. This points out the need for task-specific evaluation metrics, which take into account elements that shape user experience.

• **Incorporation of elements from the visual modality**. Since subtitling is highly multimodal and intersemiotic, ignoring elements from the visual modality can result to errors. Some features from the visual modality are already integrated in many (non-MT) tools, e.g. marking of shot changes. Another useful feature could be the recognition of on-screen text.

6 Conclusions

In this work we presented findings on subtitlers' user experience and perspectives when post-editing automatically generated subtitles, based on a survey questionnaire. Subtitlers' experience was marked as neutral to positive. Thematic analysis of the open questions showed that the main issues of automatic subtitling stem from failures in speech recognition and pre-processing, which result in error propagation, translations out of context, inaccuracies in auto-spotting and suboptimal segmentation. However, subtitlers acknowledge the positive sides of the technology, which are speed and reduction of effort, especially related to the technical aspects, as well as the provision of useful suggestions. We conclude that, despite current limitations, automatic subtitling tools can be beneficial for subtitlers, as long as improvements consider subtitlers' opinions, and ethical and professional standards are respected. We expect that as automatic subtitling tools mushroom, larger studies will be needed to explore different variables and monitor the progress in automatic subtitling.

Acknowledgements

We kindly thank all the subtitlers who took part in the survey, and Anna Matamala and María Eugenia Larreina Morales for their useful feedback on questionnaire analysis.

References

Audiovisual Translators Europe. 2021. AVTE Machine Translation Manifesto. https://avteurope.eu/wp-content/uploads/2021/09/Machine-Translation-Manifesto_ENG.pdf. Last accessed: 31/03/2022.

Bentivogli, Luisa, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2873–2887, Online, August. Association for Computational Linguistics.

Biçici, Ergun and Deniz Yuret. 2011. Instance Selection for Machine Translation using Feature Decay Algorithms. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 272–283, Edinburgh. Association for Computational Linguistics.

Bolaños-García-Escribano, Alejandro, Jorge Díaz-Cintas, and Serenella Massidda. 2021. Latest advancements in audiovisual translation education. *The Interpreter and Translator Trainer*, 15(1):1–12.

Braun, Virginia and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Buet, François and François Yvon. 2021. Toward Genre Adapted Closed Captioning. In *Interspeech 2021*, pages 4403–4407, Brno (virtual), Czech Republic, August. ISCA.

Bundgaard, Kristine. 2017. Translator Attitudes towards Translator-Computer Interaction - Findings from a Workplace Study. *HERMES - Journal of Language and Communication in Business*, 56:125–144.

Burchardt, Aljoscha, Arle Lommel, Lindsay Bywood, Kim Harris, and Maja Popović. 2016. Machine translation quality in an audiovisual context. *Target*, 28(2):206–221.

Bywood, Lindsay, Panayota Georgakopoulou, and Thierry Etchegoyhen. 2017. Embracing the threat: machine translation as a solution for subtitling. *Perspectives*, 25(3):492–508.

C. M. de Sousa, Sheila, Wilker Aziz, and Lucia Specia. 2011. Assessing the Post-Editing Effort for Automatic and Semi-Automatic Translations of DVD Subtitles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 97–103, Hissar, Bulgaria. Association for Computational Linguistics.

Cadwell, Patrick, Sharon O'Brien, and Carlos S. C. Teixeira. 2018. Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, 26(3):301–321.

Carroll, Mary and Jan Ivarsson. 1998. *Code of Good Subtitling Practice*. Simrishamn: TransEdit.

Cherry, Colin, Naveen Arivazhagan, Dirk Padfield, and Maxim Krikun. 2021. Subtitle Translation as Markup Translation. In *Proceedings of Interspeech 2021*, pages 2237–2241.

Etchegoyhen, Thierry, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine Translation for Subtitling: A Large-Scale Evaluation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 46–53, May.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, pages 1–48.

Gaido, Marco, Matteo Negri, Mauro Cettolo, and Marco Turchi. 2021. Beyond voice activity detection: Hybrid audio segmentation for direct speech translation. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing*, pages 55–62, Trento, Italy. Association for Computational Linguistics.

Guerberof-Arenas, Ana and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*.

Guerberof-Arenas, Ana. 2013. What do professional translators think about post-editing? *JoSTrans - The journal of specialised translation*, 19.

Koponen, Maarit, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020a. MT for subtitling: Investigating professional translators' user experience and feedback. In *Proceedings of 1st Workshop on Post-Editing in Modern-Day Translation*, pages 79–92, Virtual. AMTA.

Koponen, Maarit, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020b. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124, Lisboa, Portugal, November. European Association for Machine Translation.

Lakew, Surafel Melaku, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the Output Length of Neural Machine Translation. In *Proceedings of the 16th International Workshop on Spoken Language Translation, (IWSLT)*.

Laugwitz, Bettina, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In Holzinger, Andreas, editor, *HCI and Usability for Education and Work*, pages 63–76, Berlin, Heidelberg. Springer.

Liu, Danni, Jan Niehues, and Gerasimos Spanakis. 2020. Adapting end-to-end speech recognition for readable subtitles. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 247–256. Association for Computational Linguistics.

Lopes, António, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Matusov, Evgeny, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing Neural Machine Translation for Subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy, August. Association for Computational Linguistics.

Melero, Maite, Antoni Oliver, and Toni Badia. 2006. Automatic Multilingual Subtitling in the eTITLE Project. In *Proceedings of ASLIB Translating and the Computer 28*, November.

Moorkens, Joss and Sharon O'Brien. 2017. Assessing User Interface Needs of Post-Editors of Machine Translation. *Human Issues in Translation Technology: The IATIS Yearbook*, pages 109–130.

Nikolic, Kristijan and Lindsay Bywood. 2021. Audiovisual Translation: The Road Ahead. *Journal of Audiovisual Translation*, 4(1):50–70, Apr.

O'Hagan, Minako. 2003. Can language technology respond to the subtitler's dilemma? - a preliminary study. In *Proceedings of the 25th International Conference on Translation and the Computer*.

Popowich, Fred, Paul McFetridge, Davide Turcato, and Janine Toole. 2000. Machine Translation of Closed Captions. *Machine Translation*, pages 311–341.

Taylor, Christopher. 2016. The multimodal approach in audiovisual translation. *Target*, 2(28), December.

Tsiamas, Ioannis, Gerard I Gállego, José AR Fonollosa, and Marta R Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. *arXiv e-prints*, pages arXiv–2202.

Tuominen, Tiina. 2018. Multi-method research - reception in context. In Giovanni, Elena Di and Yves Gambier, editors, *Reception Studies and Audiovisual Translation*, volume 141, pages 69–90. BTL.

Vieira, Lucas Nunes. 2020. Automation anxiety and translators. *Translation Studies*, 13(1):1–21.

Volk, Martin, Rico Sennrich, Christian Hardmeier, and Frida Tidström. 2010. Machine Translation of TV Subtitles for Large Scale Production. In Zhechev, Ventsislav, editor, *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC'10)*, pages 53–62, Denver.