MT for Subtitling: Investigating professional translators' user experience and feedback

Maarit Koponen maarit.koponen@helsinki.fi
Department of Digital Humanities, HELDIG, University of Helsinki, Helsinki, Finland

Umut Sulubacak umut.sulubacak@helsinki.fi
Department of Digital Humanities, HELDIG, University of Helsinki, Helsinki, Finland

Kaisa Vitikainen kaisa.vitikainen@yle.fi
Yleisradio Oy, Helsinki, Finland

Jörg Tiedemann jorg.tiedemann@helsinki.fi
Department of Digital Humanities, HELDIG, University of Helsinki, Helsinki, Finland

Abstract

This paper presents a study of machine translation and post-editing in the field of audiovisual translation. We analyse user experience data collected from post-editing tasks completed by twelve translators in four language pairs. We also present feedback provided by the translators in semi-structured interviews. The results of the user experience survey and thematic analysis of interviews shows that the translators' impression of post-editing subtitles was on average neutral to somewhat negative, with segmentation and timing of subtitles identified as a key factor. Finally, we discuss the implications of the issues arising from the user experience survey and interviews for the future development of automatic subtitle translation.

1 Introduction

Developments in translation technology and machine translation (MT), particularly the quality improvements achieved by neural machine translation (NMT) in recent years, have led to MT increasingly becoming part of the modern-day translators' toolkit. Although post-editing (PE), where MT is used to produce a raw translation output which is then checked and corrected by a translator, has increased in many areas of translation, its use remains uncommon in audiovisual translation (AVT). AVT approaches include dubbing, voice-overs and subtitling for the purpose of making AV content accessible to audiences with no or limited understanding of the language of the original content. Different approaches are used to varying degrees depending on the type of content (e.g. voice-overs are common for documentaries) and region (e.g. subtitling is the predominant practice in Northern European countries).

As studies and practical experience have shown potential for PE in increasing productivity in other forms of translation, interest in implementing MT tools and PE workflows has also grown in the AV field. Studies have explored the use of MTPE subtitle translation with some promising although mixed results regarding effect on productivity (e.g. Bywood et al., 2017). When exploring the usability of such tools, however, productivity measurement is only one aspect. As Etchegoyhen et al. (2014) argue, subjective feedback from translators is equally important, as it provides insight into the actual user experience and necessary improvements.

This paper presents a pilot study investigating the usability of MT and PE in the subtitling

workflow from the perspective of the prospective users. Twelve professional subtitle translators working in four language pairs (Finnish↔Swedish and Finnish↔English) subtitled short video clips by post-editing MT output. We analyse feedback collected with a user experience questionnaire and semi-structured interviews for positive and negative evaluations of the PE experience and improvement suggestions. We start with an overview of related work on MT and PE in the subtitling context and work on user feedback (Section 2). After describing our approach to automatic subtitle translation (Section 3), and the subtitle PE experiment (Section 4), we present the questionnaire and interview analyses (Section 5), followed by discussion of the observations and our ongoing work based on these analyses.

2 Related work

2.1 Subtitling, MT and PE

Subtitle translation differs from translating purely textual material in that the source text consists of the spoken audio, together with the visual mode, while the target text is a written representation of translated speech. Due to technical limitations like the number of characters within a subtitle frame and the time each subtitle remains visible, paraphrasing and condensation are typical features of subtitle translation (see e.g. Pedersen, 2017). The work of subtitle translators may involve "first translation", where they translate from the source audio and determine the segmentation and timing of the subtitle frames ("spotting"), or translation with subtitle templates, where the source text consists of pre-existing intralingual subtitles in the source language or sometimes interlingual subtitles in a pivot language (often English) with set subtitle segmentation and timing (Nikolić, 2015).

To date, the use of MT and PE for subtitling has been less common in AVT than other translation fields. Explanations for this may include the characteristics of subtitle translation, which pose challenges for MT, and also the difficulty of integrating current NMT systems to subtitle translation workflows (Matusov et al., 2019). MT for movie and TV subtitling has been tested in some language pairs since the early 2000s (Melero et al., 2006; Volk et al., 2010; de Sousa et al., 2011) with suggestions that PE may increase productivity also in this context.

A subtitle-oriented statistical MT system and PE platform was developed by the SUMAT project, and tested in a user evaluation involving several language pairs and 19 professional subtitle translators (Etchegoyhen et al., 2014; Bywood et al., 2017). In a study comparing task time for translation from scratch and MTPE, Bywood et al. (2017) report that MTPE increased the translators productivity; however, the results varied for different translators, language pairs and content types. More recently, Matusov et al. (2019) tested an NMT system customised for subtitles using parallel subtitle corpora from OpenSubtitles, GlobalVoices and TED talks and reported productivity increases for MTPE in a study involving two translators.

So far, work has focused on the use of intralingual subtitles as the source text for MT, but a recent paper by Karakanta et al. (2020a) explores an end-to-end spoken language translation system for subtitling. No user evaluation of the system is reported, although Karakanta et al. (2020a) note that based on automatic evaluation against "gold standard" human subtitles the MT quality appears satisfactory. Karakanta et al. (2020b) also investigate annotating subtitle corpora for segment breaks and propose an approach for segmenting sentences into subtitles conforming to length constraints.

2.2 Studies on user experience/feedback from translators

Subjective feedback is invaluable for providing insight into tools and workflows that affect the actual work of the prospective users, and revealing issues that would not be evident from the translations or process data (see Bundgaard, 2017). Various studies have investigated professional translators' experience with and perceptions of MT and PE with questionnaires and

interviews. Analyses have reported mixed experiences: while translators sometimes find MT helpful, for example by providing useful terminology and making their work faster, other times PE may be even slower than translation from scratch. Whether working with technical or literary texts, translators often express concerns about MT affecting the final translation quality as well as their (cognitive) processes because the output can potentially mislead the translator or limit their creativity (e.g. Guerberof Arenas, 2013; Bundgaard, 2017; Moorkens et al., 2018).

Translator feedback on MT and PE in the context of AV translation was collected and analysed in the user evaluations of the SUMAT project (Etchegoyhen et al., 2014; Bywood et al., 2017). Etchegoyhen et al. (2014) describe a questionnaire used in the second evaluation round, where 19 translators carried out PE tasks in several language pairs and rated their impression of the PE process rather negatively overall (average 2.37 on a 5-point scale). Based on translator feedback, Etchegoyhen et al. (2014) identified improving MT quality to reduce cognitive load, improving quality estimation and filtering MT segments, and improving user interfaces for PE of MT subtitles as key issues for increasing usability.

Matusov et al. (2019) report a user experiment with two translators who both subtitled two programmes (a documentary and a sitcom) partly from scratch and partly with two different MT outputs. The translators rated their impression of the PE experience on average "fair" (3 on a 5-point scale) for the subtitle optimised system. The translator feedback noted useful terminology as one of the main reasons they would consider using MT in their work, but also expressed concerns about incorrect or unusual translations in the MT affecting the quality of the final translation (Matusov et al., 2019).

The study reported in this paper builds upon these analyses by collecting feedback on MT and PE for subtitling from professional subtitle translators. We aim to investigate the translators' impressions of PE more closely by introducing a more detailed user experience questionnaire where they rate different aspects of the process (see Section 4.3).

3 Automatic subtitle translation

Machine translation for subtitles requires some special treatment that we will discuss in this section. In particular, we consider models with extended context, which we will call *document-level translation models* and special tools that align translations with subtitle frames to be shown on screen. First, we briefly present the datasets and models before discussing frame alignment as a post-processing step.

3.1 Datasets and MT models

Our MT models are trained on a mix of diverse data sets[1] taken from OPUS.[2] Altogether, this includes over 30 million translation units for Finnish↔Swedish and about 44 million units for Finnish↔English. We follow the common practice in MT development to include as much data as possible even when coming from very different domains. However, the largest proportion of the training examples comes from a large collection of movie and TV show subtitles (the OpenSubtitles v2018 dataset) constituting almost half of the Finnish↔Swedish data and over 65% of the Finnish↔English data. This is certainly an advantage for our task and, hence, we expect a rather good domain-fit of our models.

We train both sentence-level and document-level models based on the Transformer architecture (Vaswani et al., 2017), the current state of the art in NMT. In particular, we apply the implementation from the MarianNMT toolkit (Junczys-Dowmunt et al., 2018), a production-ready software with fast training and decoding tools. The architecture refers to a 6-layered

[1] OPUS corpora used: bible-uedin, DGT, EMEA, EUbookshop, EUconst, Europarl, Finlex, fiskmo, GNOME, infopankki, JRC-Acquis, KDE4, MultiParaCrawl, OpenSubtitles, PHP, QED, Tatoeba, TildeMODEL, Ubuntu, wikimedia

[2] `http://opus.nlpl.eu`

network in both the encoder and decoder with 8 self-attention heads per layer. Recommended features like label smoothing and dropout are enabled and we use tied embeddings and a shared vocabulary. SentencePiece (Kudo and Richardson, 2018) is used for tokenisation and subword segmentation with models independently trained for source and target language. The shared vocabulary is set to a size of 65,000 with equal proportions in each language.

The document-level models refer to *concatenative models* proposed by Tiedemann and Scherrer (2017) and Junczys-Dowmunt (2019) using units of a maximum length of 100 tokens and special tokens for marking sentence boundaries. We observed that 100 tokens typically covers a substantial amount of contextual information in subtitles where sentences and sentence fragments are often very short. About 3.3 million pseudo-documents are created in a sequential way without overlaps for Finnish↔Swedish and 4.7 million pseudo-documents for Finnish↔English, corresponding to roughly 9 sentences per document on average.

The same kind of chunking needs to be done during test time. Sentence-level models are translated in the usual way. Note, however, that subtitles need to be pre-processed in a proper way in order to extract proper textual units that correspond to complete sentences to be translated. This involves merging fragments that run across subtitle frames and splitting frames in other cases.

We apply all our models to a dedicated test set taken from a larger set of subtitles from public broadcasts with audio in Finnish, Swedish or English. For this, intralingual subtitles (subtitles in the language of the original audio) are aligned with interlingual subtitles of the same programme in another language. The test set was carefully checked and non-corresponding segments are removed. Note that interlingual subtitles are produced independently from intralingual ones and, therefore, do not refer to direct translations of one another. Subtitles for the hard-of-hearing are also included but in a separate subset.

benchmark	**sentence-level**		**document-level**	
	BLEU	**chrF$_2$**	**BLEU**	**chrF$_2$**
fi→sv	18.8	0.443	19.3	0.451
sv→fi	15.7	0.449	16.8	0.462
fi→en	21.5	0.458	23.6	0.472
en→fi	16.0	0.444	17.1	0.454

Table 1: Comparison of BLEU and chrF$_2$ scores on the benchmark test set for the sentence-level and document-level systems in the language pairs Finnish→Swedish, Swedish→Finnish, Finnish→English, and English→Finnish.

Translation results for different subsets (scores calculated for spans of all subtitles within a video) are listed in Table 1. Evaluation of document-level translation required one additional step of aligning the automatically generated translations with corresponding reference translations. For this, we apply standard sentence alignment algorithms implemented in hunalign (Varga et al., 2005) using the re-alignment flag to enable lexical matching that ought to be very beneficial in this monolingual alignment task. Note that the automatic alignment may have negative effects on the final BLEU scores further supporting the strong result achieved by the document-level models compared to sentence-level ones according to the automatic evaluation. The scores indicate a consistent gain in using document-level information in both language pairs and all translation directions. Later, in Section 5, we will see, however, that the encouraging result does not hold in the manual assessment, which is most probably due to problems in segmentation and time frame alignment that we will discuss in the section below.

3.2 Subtitle frame alignment

One of the crucial steps in subtitle translation is the assignment to appropriate time slots. Our approach is to map translations back into the frames defined in the original source language subtitles assuming that they can fit in a similar way as the source language text was segmented. Those subtitle frames may include multiple sentences and sentences may stretch over several frames. Sentence extraction from the original subtitles is done with the techniques proposed by Tiedemann (2008). Time allocation of the translated sentences is implemented as yet another alignment algorithm.

Subtitles converted to sentence-level segments in XML:

```
  <s id="13">
    <time id="T16S" value="00:01:05,960" />
We have to make readmission agreements with
other countries, –
    <time id="T16E" value="00:01:12,360" />
    <time id="T17S" value="00:01:12,440" />
so that they would be willing.
  </s>
  <s id="14">
We have to cooperate closely.
    <time id="T17E" value="00:01:17,440" />
  </s>
```

Mapped back to subtitle frames after translation:

```
16
00:01:05,960 --> 00:01:12,360
Meidän on tehtävä
takaisinottosopimuksia muiden maiden kanssa,

17
00:01:12,440 --> 00:01:17,440
jotta ne olisivat halukkaita.
Meidän on tehtävä tiivistä yhteistyötä.
```

Figure 1: Pre- and post-processing of subtitle data before and after translation. Sentences may run over several subtitle frames and multiple sentences and sentence fragments can also appear in the same time frame. The translation comes from a document-level model.

Once again, we apply a length-based sentence alignment model to map translations to the given time slots in the source language frames. In contrast to standard bitext alignment we are now interested in 1-to-n alignments only in which each existing subtitle frame needs to be filled with one or more segments coming from the automatically generated translations. For the target language segmentation we consider simple heuristics for splitting sentences into clauses by breaking strings that are separated by punctuation plus space characters. Resulting sequences that exceed a certain length threshold are further split on space characters closest to the center of the string. After that, we apply the famous Gale & Church algorithm (Gale and Church, 1993) to optimise the global alignment between source segments (original subtitle frame data) and target segments with adjusted parameters referring to our specific task: (1) We apply a uniform prior over alignment types as there is no strong preference for frame-to-clause alignment in our case. (2) We define alignment types to include one-to-x units only with x ranging from one to four. (3) We introduce extra costs to discourage frame boundaries within running sentences and assignments that violate length constraints. Figure 1 shows an example outcome of the procedure.

Finally, we also apply simple heuristics to insert line breaks making them conform to length and formatting constraints. During the manual assessment, we found out that this segmentation and the introduction of length violation costs caused severe damage to the time slot assignment pointing out the importance of proper optimisations of those steps. The implementation of our frame alignment algorithm is available as an open source package.[3]

4 Subtitle post-editing experiment and collecting user feedback

The study described in this paper is a part of a research project involving MT and other technologies as a tool for managing and processing AV material. The purpose of this study was

[3] https://github.com/Helsinki-NLP/subalign

to investigate the usability of MT for interlingual subtitling in an experiment carried out in November–December 2019 with professional subtitle translators using MT and PE to subtitle short video clips. The experiment involved recording process data with keylogging software (Inputlog, see Leijten and Van Waes, 2013). The translators' subjective evaluations of the usability of MTPE for subtitling were collected with a questionnaire focused on the user experience and semi-structured interviews. In this paper, we focus on these subjective evaluations and the professional translators' user experience of MT and PE for subtitling.

4.1 Participants and the subtitling workflow context

The experiments were carried out at the Finnish public broadcasting company Yle which produces and broadcasts AV content on television and an online streaming service. The company employs in-house translators and outsources some translation work. Subtitling is the most common approach to AVT in this company and Finland in general. Subtitle templates are generally not used by Yle, rather, the translators' normal workflow involves first translation from audio and spotting the subtitles manually. The translators follow quality recommendations which specify, for example, technical recommendations like number of characters in subtitle frames, minimum and maximum duration of subtitle frames on screen and maximum reading speed, as well as linguistic features. National guidelines for subtitle translation published in 2020[4] reflect the practices already in place at the broadcasting company.

The subtitling tasks were carried out in four language pairs: Finnish→Swedish, Swedish→Finnish, Finnish→English and English→Finnish. Twelve translators (three per language pair) participated in the experiment: eight in-house translators and four freelancers with between 4 and 30 years of professional experience as subtitle translators in the relevant language pair. Two participants stated they had experimented with using MT for subtitling prior to this test, and seven others had used MT for other purposes.

4.2 Materials and subtitling tasks

Video clips to be subtitled were selected from datasets representing two content types: EU election debates (unscripted dialogue between multiple participants) and lifestyle or cultural programmes (semi-scripted dialogue or monologue by programme hosts on various topics e.g. movies, food and drink). Each clip was selected so that it (1) formed a coherent, self-contained section of the program as a whole; (2) was approximately 3 minutes long; and (3) contained approximately 30–35 intralingual subtitles. The length and number of clips was limited due to the limited availability of participants for the experiments. Some clips consisted of complete programmes of suitable length, while others were cut from longer programmes ensuring that they formed a coherent, self-contained section. Human-generated intralingual subtitles in the source language were translated with two different MT models, and aligned to SRT files using the subtitle segmentation and timing from the intralingual subtitles as detailed in Section 3.1.

The subtitling tasks were carried out using the subtitlers' preferred subtitling software (Wincaps Q4 or Spot). An external monitor and keyboard were provided, and the subtitlers had access to the internet as well as terminology and other resources normally used in their work. The participants were instructed to create subtitles that would be acceptable for broadcasting, and to follow their normal working processes, but to not spend excessive time on "polishing" any given wording or on researching information. No explicit time limit was given, rather, the participants were instructed to work at their own pace.

Each participant carried out six tasks: MTPE for four clips (two clips with sentence-level MT output and two clips with document-level MT output), and translation from scratch for two clips, with spoken audio as source and manual spotting. To mitigate potential differences related

[4]Currently available in Finnish and Swedish: `http://www.av-kaantajat.fi/Laatusuositukset/`.

to difficulty of each clip and facilitation effect, the clips and MT outputs were rotated so that each clip was subtitled with no MT output, with sentence-level MT output and with document-level MT output by a different participant, and task order was varied. An experimenter was present to set up each task, assist with any potential technical issues and conduct the post-task interview, but did not interact with the participants during the tasks.

4.3 User Experience Questionnaire

An online form was used to collect subjective evaluations of the usability of the MT output for PE. The questionnaire was based on the User Experience Questionnaire (UEQ) developed by Laugwitz et al. (2008) for end-user evaluation of software products. The objective of the UEQ is to provide users with a "simple and immediate way to express feelings, impressions and attitudes" toward the product and thereby elicit quick but comprehensive assessments of user experience (Laugwitz et al., 2008, 64). It consists of scalar ratings of opposing adjective pairs (e.g. *practical/impractical*) intended to measure both classic usability aspects and user experience aspects. The adjective pairs are shown on an scale of 1–7, with positive and negative adjectives alternating on the left/right.

Because the focus of our study was on the participants' experience of MTPE rather than subtitling software, a modified version of the UEQ was created to focus on the participant's experience of PE as a process. Adjective pairs focusing on the attractiveness or usability of the software interface were omitted, and some adjective pairs were added to elicit responses more focused on PE. The final questionnaire was provided to the participants in Finnish, and contained the following 13 adjective pairs[5], preceded by the words *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*. For analysis, we processed the scores using the formulae in the UEQ Data Analysis Tools (version 7)[6] 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.

In addition to the PE experience, we included Likert-scale assessments for the automatic spotting and segmentation of subtitle frames (1 poor – 7 good) and the effort needed to correct them (1 easy – 7 a lot of effort). Short open questions were included for more specific comments regarding the MT output, subtitle spotting and segmentation.

4.4 Semi-structured interviews

After completing all PE tasks, a brief semi-structured interview was also carried out to collect more detailed feedback on each participant's experience, features affecting the process and usability, and possible suggestions for future development and improvements. In the interview, the participants were first asked for their overall impression of the PE tasks was, what features of the MT output affected that impression experience and whether they observed differences between the outputs. They were then asked to describe their normal subtitling process and how the MT output affected that process. Finally, the participants were asked whether they would consider using MT as a tool in their own work and what kind of improvements would be needed.

The interviews were transcribed and anonymised, and thematic analysis (see e.g. Matthews and Ross, 2010) was carried out using the analysis software Atlas.ti. The interview responses were analysed for positive and negative comments and specific issues raised by the participants,

[5]The Finnish translations provided in version 8 of UEQ were not yet available at the time of our experiment. The Finnish adjective pairs were created by the authors, and we provide here our back-translations into English.

[6]https://www.ueq-online.org/

such as features impacting quality and usability or suggestions for improvement. In some cases, the participant's statement was not explicitly positive or negative, but rather consisted of a neutral, generic observation or a mixed evaluation, such as a comment (sv-fi, participant A) that the MT was "sometimes surprisingly good but sometimes surprisingly bad". Such cases were labelled as mixed/neutral.

5 Results

5.1 Evaluations of User Experience

Figure 2 shows the UEQ scores for each adjective pair averaged over all participants and all clips in each language pair comparing the two MT outputs (sentence-level model and document-level model). On average, the participants appeared to describe their MTPE experience in neutral terms (values between -0.8 and +0.8). Averaged across all participants, the most negative reactions were seen for *labourious/effortless* and *limiting/creative*, although these did not cross the -0.8 threshold. No clear differences emerged between the two different MT outputs, although the participants appeared to have a slight preference for the sentence-level MT output. Similarly, no clear difference was observed for the two programme types. Overall scores for lifestyle/cultural clips were slightly higher, except in Swedish→Finnish, where the election debate clips received slightly higher scores.

Interestingly, the participants' experiences appeared to differ in different language pairs. In particular, the participants working with English→Finnish evaluated nearly all adjective pairs negatively; only *stressful/relaxed* and *complicated/simple* show neutral averages in this language pair. Responses for Swedish→Finnish were more neutral, although tending toward negative. For Finnish→Swedish, evaluations were generally neutral, except *difficult/easy* and *complicated/simple*, where averages for the document-level output crossed the 0.8 threshold to positive evaluation. Finally, for Finnish→English clearly negative scores were seen only for the adjective pair *limiting/creative*, and the sentence-level output reaches positive averages for *difficult/easy*, *stressful/relaxed*, *inefficient/efficient*, *complicated/simple* and *demotivating/motivating*.

Spotting and segmentation of subtitle frames was generally assessed as poor, and problems appeared to have been more common in the document-level output. Correcting spotting and segmentation, however, was mostly characterised as neutral or easy. The participants working with English→Finnish assessed spotting/segmentation as particularly poor and difficult to correct, which may have affected their general impression of the PE process as a whole.

5.2 Analysis of positive and negative statements in user interviews

Table 2 shows the numbers of positive, negative and mixed/neutral statements identified. Of the total 143 statements, 55% (79) were classified as negative. Positive statements accounted for 29% (42) and mixed/neutral statements for 15% (22). No differences were observed between the two MT outputs. most statements characterised MT in general without reference to specific output. In cases where a specific output was identifiable, the numbers of positive, negative and mixed statements were roughly equal between the two outputs. However, some differences can be seen between language pairs. The proportion of negative statements is higher in Swedish→Finnish and English→Finnish than in the other two language pairs. Finnish→English translators have the highest number of positive statements, and Finnish→Swedish translators the highest proportion of mixed/neutral statements.

A more detailed analysis was also conducted to identify the specific issue discussed in negative and positive statements. Most common issues involved spotting/segmentation of the subtitles, MT output quality, and the effect of MT and PE on the translator's workflow and processes. Some statements also concerned other issues like the clips and their subject matter.

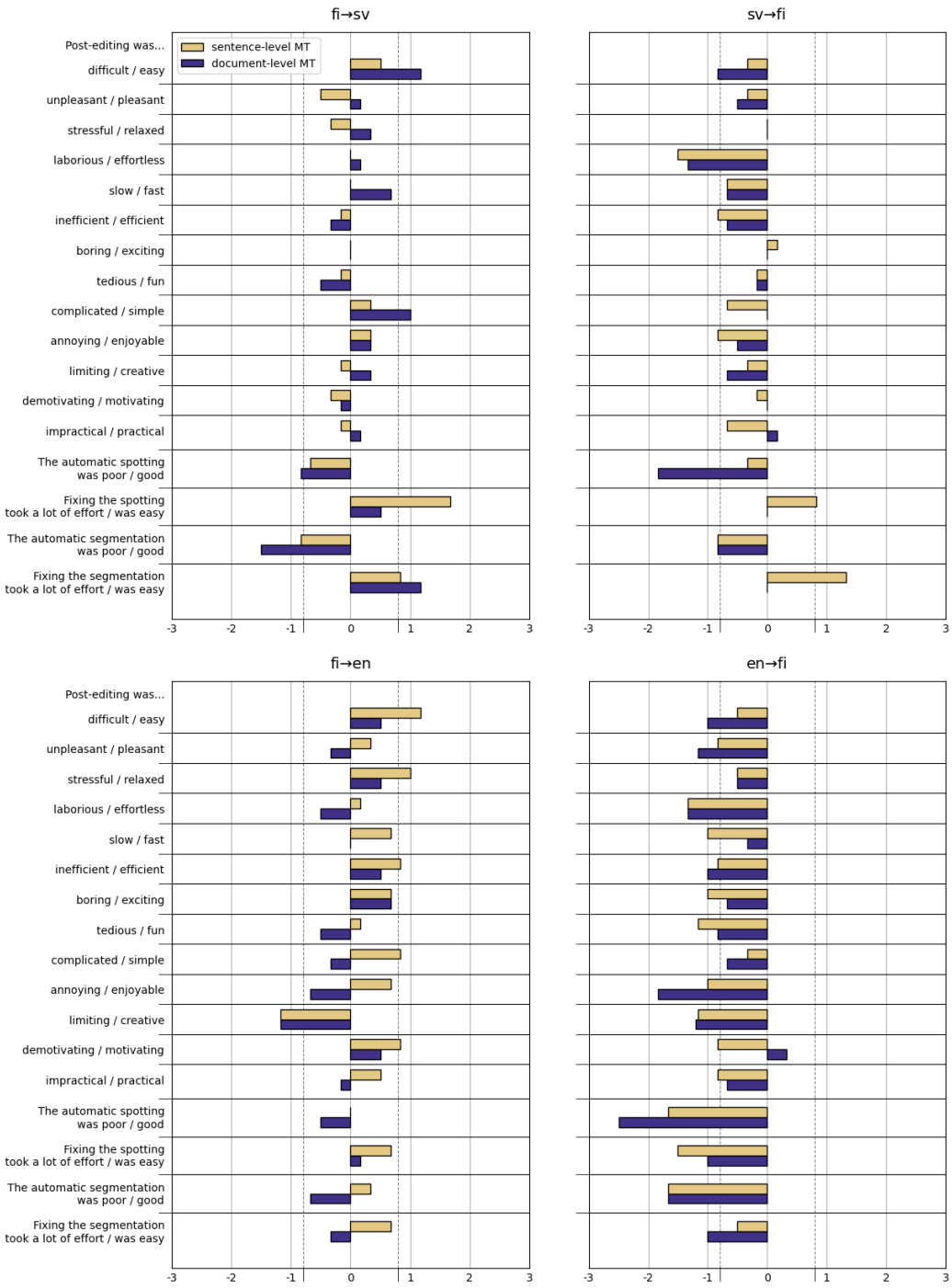

Figure 2: Average user experience scores comparing post-editing of sentence-level and document-level MT output in the language pairs Finnish→Swedish, Swedish→Finnish, Finnish→English and English→Finnish. Dashed lines represent the range [-0.8, +0.8], defined as neutral evaluations in the UEQ Data Analysis Tool.

Statement type	en→fi	fi→en	fi→sv	sv→fi	Total
Positive	13	16	8	5	42
Negative	23	19	14	23	79
Mixed/neutral	2	3	10	7	22
Total	38	38	32	35	143

Table 2: Positive, negative and mixed/neutral statements in the translator interviews

Most negative statements (33 out of 79) concerned the spotting or segmentation of subtitle frames, which all 12 participants commented on negatively. Specific problems involved subtitles being out of sync with the audio, and cases where a sentence had been incorrectly split into two (or more) segments. Two participants felt that the MT output tried to pack "too much" into a subtitle frame and that the machine was not able to condense the translation. Although the translations were created based on intralingual subtitles, which often already involve some condensation compared to the audio, this may suggest further differences between source and target languages. Three statements regarding the spotting were mixed/neutral, and the only two positive statements qualified spotting as "better" in some clips.

MT output quality received 30 negative mentions. Specific issues included lexical errors like mistranslated words or "odd" word choices (8 statements) and accuracy errors involving longer passages (5 statements), as well as fluency issues like ungrammatical or unidiomatic structures (6 statements). Two participants also noted omissions (words or longer passages) in the MT output. The remaining negative statements referred to MT output in general, without naming specific issues. On the other hand, the participants made 23 positive statements concerning MT quality. Specific comments referred to useful terminology and other lexical choices (9 statements) and fluency of the output (3 statements), while 11 positive statements involved general characterisations of the output as good or useful. Additionally, 13 mixed/neutral statements were made involving MT output quality in general terms.

The effect of MT and PE on the subtitling process was mentioned in 42 statements, which were mostly negative. In 15 statements, the participants commented that using MT and PE seemed to involve more effort than translation from scratch and to reduce productivity. A positive effect on productivity was mentioned in 3 statements, and 9 statements characterised the effect as mixed, sometimes reducing but sometimes increasing effort. Negative comments regarding effects also included an impression of being limited by the MT (8 statements) and potentially lower quality of the final translation (5 statements). Finally, 12 statements were made characterising the overall PE experience positively, while 5 statements described the experience negatively, and 8 in mixed/neutral terms.

5.3 User feedback for improvements

In total 28 suggestions involving development and improvement were identified in the transcripts. The most commonly mentioned improvement need was spotting/segmentation of the subtitles (8 statements). Two participants mentioned segmentation according to speaker changes as particularly useful. On the other hand, two participants would have preferred to see the MT output separately without segmentation, and one wished to see automatic speech recognition output of the original audio. Other specific issues mentioned involved need for condensing the MT output for subtitles (2 statements), improving cohesion, genre adaptation and punctuation in subtitles. Two participants mentioned the multimodal nature of AVT, one remarking that the machine is not able to take the visual aspect into account and the other wondering whether MT could use visual information.

Integration of functionalities other than MT into the subtitling software was also mentioned by some participants. Some type of terminology tool integration was mentioned in 4 statements. Some wished for a tool like a translation memory (4 statements), where one or more translators could add their own material and see how things were translated previously.

Of the participants, four would consider using MT for subtitling, although all would like to see some improvements in quality, while two participants stated they could not see themselves using MT as a tool at all. The other six gave a more mixed answer, stating that they could see MT and PE suitable for some situations but not others, for example, depending on the type of programme and subject matter. Some considered MT most useful for unfamiliar content as a terminology aid. In contrast, others would only use MT with subject matter they were already familiar with, to make sure to notice possible errors. With regard to genre, some stated MT seemed more useful for the election debates, with more formal speech, while others considered it more suitable for "simpler", less formal language in some of the lifestyle clips.

6 Discussion and ongoing work

The subjective evaluations offer valuable insight into the user experience of MTPE for subtitling. Our participants did not find PE particularly difficult or complicated, but they tended to characterise it negatively as limiting and annoying in the questionnaire, and these themes are further present in the interviews. The translators' feelings of MT limiting their creativity are similar to findings in studies addressing literary translation (Moorkens et al., 2018) as well as localisation (Guerberof Arenas, 2013). Translators in other studies have similarly referred to being "trapped by MT" (Bundgaard, 2017) and expressed concerns of a detrimental effect on the quality of the final translation (Moorkens et al., 2018; Matusov et al., 2019).

Both the questionnaire and the interviews point to problems in the MT subtitle alignment. Although the frame alignment (see Section 3.2) produces subtitles conforming to technical length constraints, the translators did not always find *way* the content was segmented acceptable. In some files, omissions or repetition of content in the MT also caused misalignments. The overall assessment of user experience also appears more negative in the language pairs where the participants rated the timing and segmentation poorest. This suggests that alignment problems may have affected the overall experience in addition to the MT output quality. One participant explicitly stated that dealing with off-sync subtitles probably led them to make also linguistic changes that may have been unnecessary.

In the interviews, most participants did not think MTPE increased productivity, some even felt the opposite. Similar observations have again been made in other studies; Etchegoyhen et al. (2014), for example, discuss how the increased cognitive load of dealing with MT output is a significant part of productivity. Although a detailed discussion of the process data is not within the scope of this paper, some parallels can be seen in our productivity measurements. On average, task times for MTPE were slightly faster than for translation from scratch, although considerable variation was observed between different files and participants. Five out of twelve participants were in fact slower when post-editing, concurring with the translators' mixed experience. For a more detailed analysis of the productivity metrics, see Koponen et al. (2020).

Some care is needed when interpreting the results. Firstly, it is important to note that the participants in this study did not have prior experience with MT for subtitling (only two had previously tested it). Their responses may therefore be affected by the unfamiliarity of the task, which some participants mentioned in the interviews (see also similar observations by Bywood et al., 2017). Secondly, the participants are used to doing first translation from audio instead of working with subtitle templates. Since the MT outputs were created using intralingual subtitles as the source text, rather than the spoken language, the participants may additionally have been affected by the intralingual subtitler's choices regarding spotting, paraphrasing and con-

densation. These may then have been perceived as issues in the MT output, such as omissions. More detailed communication regarding how the subtitles had been automatically generated could have clarified this issue for the participants. Finally, to allow the participants to follow their normal subtitling processes, they used their preferred subtitling software in the PE tasks. However, these tools (and AVT tools in general) are not designed for MTPE, and may therefore not be optimal for the task. This may also have affected the participants' perception of PE, and exploring AVT tools with more effective support for MTPE would be needed.

In view of our observations, it is clear that more work is needed to address the issues pointed out by the participants. It seems relevant to also compare experiences in a contrastive MTPE setting based on automatic AV transcriptions, in order to neutralise creative constraints imposed by subtitling choices carried over from intralingual subtitles. Following the interviews, we have made an effort to improve our MT pipeline in response to the segmentation and time frame alignment issues, and added support for machine-generated transcripts and time frames via automatic speech recognition and spotting. We have fixed some errors in our segmentation procedure for subtitle translations, and updated our heuristics to be less strict in enforcing length limits and clause breaks. Currently, our pipeline also makes use of an additional *restoration* stage as an endcap for MT pre-processing, implemented in practice as "intralingual translation" going from case- and punctuation-stripped input to fully-formatted output within the same language. The goal of this stage is to boost translation performance on automatic transcripts (where the MT is sensitive to differences in input formatting), and also of segmentation heuristics for post-processing (which are heavily dependent on punctuation in determining clause boundaries). Improvements to the MT and segmentation have been evaluated in further MTPE user tests during summer/fall 2020 with most of the same participants, and analysis of the data is still underway. Preliminary observations suggest somewhat more positive views of MT quality and segmentation, but the use of automatic transcriptions was received more negatively.

7 Conclusion

In this paper, we have presented a user evaluation of MTPE for subtitle translation based on experiments carried out by twelve professional subtitle translators in four language pairs (Finnish↔Swedish and Finnish↔English). Our analysis of data collected with a user experience questionnaire showed that, on average, translators' impression of MTPE varied from negative to neutral or mildly positive depending on language pair. Thematic analysis of interviews provided further information of the translators' experience. While translators did not consider PE particularly difficult, they tended to characterise it as limiting and somewhat annoying. Most, however, were open to using MT for at least some subtitling content. Further work on the quality of the outputs and tools is needed, and the translators' feedback provided valuable insight for this work. In both the questionnaire and interviews, the segmentation and timing of MT subtitles were identified as major issues, in addition to overall MT quality. As this paper reports our first user evaluations of MT for subtitling in specific language pairs and in a specific AVT context, definitive conclusions regarding the ultimate applicability of MTPE for subtitling naturally cannot yet be made. As work in this area continues, further studies on the user experience of subtitle translators are essential to investigate this question.

Acknowledgements

This work is part of the MeMAD project, funded by the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No 780069).

References

Bundgaard, K. (2017). Translator attitudes towards translator-computer interaction - Findings from a workplace study. *Hermes– Journal of Language and Communication in Business*, 56:125–144.

Bywood, L., Georgakopoulou, P., and Etchegoyhen, T. (2017). Embracing the threat: machine translation as a solution for subtitling. *Perspectives: Studies in Translatology*, 25(3):492–508.

de Sousa, S. C., Aziz, W., and Specia, L. (2011). Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles. In *Proceedings of RANLP 2011*, pages 97–103.

Etchegoyhen, T., Bywood, L., Fishel, M., Georgakopoulou, P., Jiang, J., Van Loenhout, G., Del Pozo, A., Maučec, M. S., Turner, A., and Volk, M. (2014). Machine translation for subtitling: A large-scale evaluation. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 46–53.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Guerberof Arenas, A. (2013). What do professional translators think about post-editing. *The Journal of Specialised Translation*, 19(19):75–95.

Junczys-Dowmunt, M. (2019). Microsoft Translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 225–233.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Karakanta, A., Negri, M., and Turchi, M. (2020a). Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 209–219, Online. Association for Computational Linguistics.

Karakanta, A., Negri, M., and Turchi, M. (2020b). MuST-cinema: a speech-to-subtitles corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3727–3734, Marseille, France. European Language Resources Association.

Koponen, M., Sulubacak, U., Vitikainen, K., and Tiedemann, J. (2020). 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. European Association for Machine Translation.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 EMNLP*, pages 66–71.

Laugwitz, B., Held, T., and Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In Holzinger, A., editor, *HCI and Usability for Education and Work. USAB 2008*, volume 5298 of *Lecture Notes in Computer Science*, pages 63–76, Berlin/Heidelberg. Springer.

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

Matthews, B. and Ross, L. (2010). *Research Methods: A Practical Guide for the Social Sciences*. Pearson Education Ltd, Edinburgh.

Matusov, E., Wilken, P., and Georgakopoulou, Y. (2019). Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation*, pages 82–93.

Melero, M., Oliver, A., and Badia, T. (2006). Automatic multilingual subtitling in the eTITLE project. In *Proceedings of Translating and the Computer 28*, pages 1–18.

Moorkens, J., Toral, A., Castilho, S., and Way, A. (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262.

Nikolić, K. (2015). The pros and cons of using templates in subtitling. In *Audiovisual Translation in a Global Context: Mapping an Ever-Changing Landscape*, pages 192–202.

Pedersen, J. (2017). The FAR model: assessing quality in interlingual subtitling. *The Journal of Specialised Translation*, 28:210–229.

Tiedemann, J. (2008). Synchronizing translated movie subtitles. In *Proceedings of LREC'08*.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third DiscoMT*, pages 82–92.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Volk, M., Sennrich, R., Hardmeier, C., and Tidström, F. (2010). Machine translation of TV subtitles for large scale production. In *Proceedings of the Second Joint EM+/CNGL Workshop*, pages 53–62.