Journal of
Audiovisual Translation

Automation in the Intralingual Subtitling Process: Exploring Productivity and User Experience

Kaisa Vitikainen✉
Yle and University of Helsinki

Maarit Koponen✉
University of Eastern Finland

Citation: Vitikainen, K. & Koponen, M. (2021). Automation in the intralingual subtitling process: Exploring productivity and user experience. *Journal of Audiovisual Translation, 4*(3), 44–65. https://doi.org/10.47476/jat.v4i3.2021.197

Editor(s): H. Silvester & T. Tuominen

Received: July 14, 2021

Accepted: October 15, 2021

Published: December 28, 2021

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

Abstract

The demand for intralingual subtitles for television and video content is increasing. In Finland, major broadcasting companies are required to provide intralingual subtitles for all or a portion of their programming in Finnish and Swedish, excluding certain live events. To meet this need, technology could offer solutions in the form of automatic speech recognition and subtitle generation. Although fully automatic subtitles may not be of sufficient quality to be accepted by the target audience, they can be a useful tool for the subtitler. This article presents research conducted as part of the MeMAD project, where automatically generated subtitles for Finnish were tested in professional workflows with four subtitlers. We discuss observations regarding the effect of automation on productivity based on experiments where participants subtitled short video clips from scratch, by respeaking and by post-editing automatically generated subtitles, as well as the subtitlers' experience based on feedback collected with questionnaires and interviews.

Key words: automatic subtitling, automatic speech recognition, intralingual subtitling, productivity, user experience.

✉ kaisa.vitikainen@yle.fi, https://orcid.org/0000-0003-2067-3969
✉ maarit.koponen@uef.fi, https://orcid.org/0000-0002-6123-5386

1. Introduction

Intralingual subtitling refers to the process of creating subtitles within the same language (Gottlieb 1997), which then appear in a television broadcast or other audiovisual content as open or closed subtitles, also called captions (Liu, 2014). Intralingual subtitles provide a valuable service for many users, including persons with hearing impairment, immigrants and elderly persons. In recent years, the need for intralingual subtitles has increased and the service is increasingly targeted by legal requirements (see e.g. Romero-Fresco 2011 for discussion). In Finland, the law requires broadcasters to provide intralingual subtitles for content in Finnish and Swedish. The requirement is 100% of content for the public broadcaster and 75% of content for major commercial broadcasters, except certain live events (Kiuru et al., 2020). At the same time, pressure for lower budgets and shorter turnaround times is increasing. To meet this need, machine learning solutions such as automatic speech recognition (ASR) could be valuable tools for both broadcasters and subtitlers.

Fully automatic subtitles are used by some broadcasters in Finland, but reaction from the target audience has been negative. For example, Kiuru et al. (2020) conducted a survey aimed at the users of intralingual subtitles, including people with hearing and speech impairments, immigrants, and the elderly. The survey was distributed by various associations representing the target groups and had 137 respondents. Kiuru et al. (2020) report that users were very critical of the quality of automatic subtitles: many respondents even found them offensive, and some said that automatic subtitles should be forbidden by law. Kuuloliitto, an advocacy group for the Deaf and the hard of hearing, has also harshly criticised fully automatic subtitles. For example, Kuuloliitto submitted a statement to the Ministry of Transport and Communications regarding the automated subtitles offered by one Finnish channel. They noted that the subtitles were inaccurate, that viewers could not distinguish speakers due to the lack of punctuation, and that many hearing-impaired viewers found the programming impossible to watch (Kuuloliitto, 2019).

The negative response indicates that fully automatic subtitles for Finnish have not yet reached sufficient quality to be accepted by the target audience, although studies by Tiittula et al. (2018) on the reception of subtitles produced by different means (respeaking, ASR directly from original audio, manual transcription) and by Kóbor-Laitinen (2021) on the reception of fully automatic subtitles among a hard of hearing audience suggest that audiences could be open to automated subtitles with quality improvements. The legislation requiring broadcasters to provide intralingual subtitles for their Finnish and Swedish content was changed in 2021 to include a clause stipulating that "the subtitling service must be implemented with high quality in such a way that the subtitles are sufficiently clear and understandable to the user."[1] Broadcasters must then consider quality as well as quantity as they produce subtitles for their content. National quality guidelines for intralingual subtitles (Vitikainen et al. 2020) were published in January 2021.

[1] "Tekstityspalvelu tulee toteuttaa laadukkaasti siten, että tekstitys on käyttäjälle riittävän selkeää ja ymmärrettävää" (Laki sähköisen viestinnän palveluista [Law on Electronic Communication Services] 211 §; translation by authors).

Although fully automatic subtitling may not be feasible, ASR can potentially be useful as a tool for the subtitler. One option is a workflow where ASR is used for subtitling by a respeaker[2] rather than directly transcribing the original speech, which is used in many countries for live broadcasts (Romero-Fresco, 2011), although ASR applications exist for non-live subtitling as well. Using ASR to generate subtitles directly from the original speech has also been tested in some previous work mainly for English and some other languages (see Álvarez et al., 2016b; Matamala et al., 2017; Tardel, 2020). However, little research has previously been done on automatic Finnish speech recognition as a tool for the subtitler. Vitikainen (2018) studied the first steps towards respeaking in Finnish, but to our knowledge no previous work has addressed automatic transcription of the original speech as part of a professional subtitling workflow.

This article aims to examine ASR combined with automatic timecoding and segmentation as a tool for intralingual subtitlers in the public service broadcasting context. We examine a workflow where these tools are used to automatically generate subtitles which are then post-edited by the subtitlers, and investigate the impact on productivity as well as the subtitlers' user experience and attitudes. This work has been conducted as part of the EU-funded MeMAD project (Methods for Managing Audiovisual Data), which explored combining machine learning with human effort in different work processes involving audiovisual materials. The use of automatic intralingual subtitling was tested in the project together with in-house subtitlers of project partner Yle (Finnish Broadcasting Company). In two rounds of evaluation experiments, conducted in 2019 and 2020, participants subtitled short video clips by post-editing automatically generated subtitles, by respeaking, or from scratch, using their usual work process. In the first experiment round, process data were collected through keylogging to compare measures of productivity. Feedback was collected through forms based on the User Experience Questionnaire (UEQ, Laugwitz et al. 2008) after each task and through interviews after each round of experiments. The general user-oriented approaches and some preliminary observations have been previously discussed in Koponen et al. (2020). In this paper, we focus on a more detailed investigation of the experiments conducted with the intralingual subtitlers.

This article is structured as follows. Section 2 presents an overview of previous work that has focused on the use of ASR in the intralingual subtitling process. In Section 3, we describe the tools and experiment set-up for the productivity tests, user experience questionnaire and interviews. Section 4 presents the findings of these experiments, followed by discussion and concluding remarks in Section 5.

2. Intralingual subtitling and automatic speech recognition

In recent years, the use of speech recognition technology has increased in various scenarios from dictation to transcription and for different types of content (Álvarez et al. 2016b). In addition to

[2] In live respeaking, the subtitler/respeaker repeats what is said, editing as needed and adding punctuation, to a speech recognition software, which generates the subtitles.

transcribing audiovisual content (see e.g. Tardel 2020), ASR has also been used to generate intralingual subtitles in some previous projects. Previous work has commonly focused on ASR for English, for example, in the MUSA project[3] (Piperidis et al., 2004), the HBB4All[4] project (Matamala et al., 2015) and the ALST project (Matamala et al., 2017). Other languages include Spanish in the APyCA project (Álvarez et al., 2010), as well as Basque, Spanish and Italian in the SAVAS project (Álvarez et al., 2016b). The content types subtitled in these studies have typically involved different types of non-fiction content, such as documentaries, news, weather forecasts and political interviews. The EU-BRIDGE[5] project also involved a use case for captioning television broadcasts, YouTube videos and lectures in various languages.

Although the use of ASR has increased in subtitling, automatically generated subtitles have limitations. High-quality speech recognition is available only for some languages, and quality is also affected by the speakers' characteristics and other features like background noise. Furthermore, the tools generally cannot automatically segment the transcript into subtitles that follow professional conventions for punctuation and reading speed (see Álvarez et al., 2016a,b). Due to these issues, the automatic output requires post-editing by a subtitler to correct errors and modify the subtitles to be suitable for broadcasting. A key question is how this use of automation affects the subtitling process.

While process study methodologies like screen recording, keystroke logging and eyetracking have increased understanding about the process of written translation, few studies have examined the audiovisual setting and the subtitling process, generally focusing on interlingual translation (Orrego-Carmona et al., 2018, pp. 151-153). Studies involving the use of technology like ASR in subtitling have also focused more on product than process (Tardel 2020). One previous study by Álvarez et al (2016b) specifically addressed temporal effort when post-editing automatically generated subtitles. Álvarez et al. (2016b) compared the productivity of five professional subtitlers when subtitling from scratch and when post-editing automatically generated intralingual subtitles, and found that for four participants productivity, in terms of subtitles per minute, increased 2% to 33%, whereas the fifth participant was slower when post-editing. Another study involving nine student subtitlers found that improvements in the automatic segmentation method increased productivity and decreased perceived effort in the post-editing process (Álvarez et al., 2016a). Other studies have also examined post-editing ASR transcription. Matamala et al. (2017) conducted an experiment with 10 professional transcribers working on English interview content, comparing manual transcription, respeaking and post-editing of ASR output. They found that post-editing was in fact the slowest of the three approaches and the participants' subjective assessments also indicated that post-editing took the most effort, was most boring and least accurate (Matamala et al., 2017). Tardel (2020) investigated different indicators of temporal, technical and cognitive effort in an experiment where 12 professional subtitlers and 13 translation students completed two intralingual transcription tasks in German as well as translation tasks. ASR post-editing was found to be slightly slower than manual

[3] http://sifnos.ilsp.gr/musa
[4] http://www.hbb4all.eu/
[5] https://www.eu-bridge.eu/captioning.html

transcription, although technical effort in terms of number of keystrokes was lower, and eyetracking data showed more frequent switches of attention between the video and ASR output compared to manual transcription (Tardel, 2020). In this paper, we aim to further explore how automatic subtitle generation affects productivity and user experience in intralingual subtitling and examine the role of subtitle segmentation and timing.

3. Post-editing experiments and collecting user feedback

This section describes experiments investigating the efficiency and usability of automatic intralingual subtitling, which were conducted as part of the MeMAD project evaluations in two rounds (2019 and 2020). Because initial experiments suggested the product maturity was high enough to be tested in production, an additional proof-of-concept period was organised between the two rounds. During the proof-of-concept period, the participants used automatic subtitles in their own work to explore whether additional experience with the post-editing process could improve the subtitlers' productivity. The experiments involved the collection of process data to examine the effect of automation on subtitling effort as well as questionnaire and interview data to evaluate usability. The following sub-sections provide information about the tools used, participants, materials, tasks, and collection of productivity data and feedback.

3.1. Tools used in the experiments

For the post-editing experiments, intralingual Finnish subtitles were generated using ASR as well as automatic segmentation and timecoding. In the first round in 2019, the subtitlers tested output from the commercial Google Speech Recognition[6] (v1 non-streaming version with default speech model) and a proprietary ASR system developed by the MeMAD project partner Lingsoft Oy (see Doukhan et al., 2020). This first round compared the users’ experiences with the different outputs. Google output was included because it was the default ASR integrated in the platform used to generate the automatic subtitles (see below), while the Lingsoft system was specifically being developed for the purposes of this project. In this first round, the subtitlers also used the respeaking software Sanelius[7] for comparison because some Yle subtitlers use offline respeaking routinely. For the second round in 2020, only the Lingsoft system was used, as the results in the first round were more promising compared to Google. Respeaking tasks were not included in the second round as it was not part of the project focus.

Segmentation and timecoding were also produced automatically using the Flow platform developed by the project partner Limecraft (see Braeckman et al., 2021). The platform was used to first produce a transcript for each clip with the integrated ASR system (Google or Lingsoft), and then to generate

[6] https://cloud.google.com/speech-to-text/
[7] http://www.sanelius.fi

timed subtitles. The timecoding and segmentation rules were set to follow Yle guidelines as closely as possible: a soft cap of 34 and a hard cap of 37 characters per line, a soft cap of 2 and a hard cap of 3 lines of text per text block (3 lines would have to be edited down to 2), 0.08 seconds gap between subtitle blocks (equivalent to 2 video frames in the PAL format, the standard at Yle), length between 2 and 5 seconds, and adding a hyphen to the end of the first block if splitting a sentence between blocks. Reading speed was not set, as the Flow platform counted reading speed as words per minute (wpm), which is not used in Finland. Finnish subtitles use characters per second (cps), with a recommended reading speed of 10-12 cps (Vitikainen et al., 2020). For the post-editing experiments, the automatically generated subtitles were exported in SRT format and imported into the subtitling tool normally used by Yle subtitlers (Wincaps Q4).

3.2. Participants

The experiments involved a total of four participants in the first round and three participants in the second round. All participants were in-house subtitlers at Yle, with between 9 to 20 years of experience as professional subtitlers. All had previously used speech recognition in some form as an aid for subtitling; three participants had used it occasionally and one participant frequently, mostly in the form of offline respeaking. Due to scheduling issues, one participant was not able to complete the last task in 2019, and one participant was not available for the second round. The 2020 experiments were therefore carried out with only three participants.

3.3. Data collection for experiments in 2019

The first experiment round was conducted in October and November 2019 in the premises of the Finnish Broadcasting Company Yle. Materials for the 2019 experiments consisted of two genres: clips from EU election debates and from a youth-oriented lifestyle programme “Onks noloo?”. Seven clips were selected overall: four from the EU debates and three from the youth programme. Each clip was approximately 3 minutes long and formed a coherent, self-contained section.

The following tasks were carried out:

- subtitling from scratch, as they would in their normal working environment, including timecoding (two tasks);
- subtitling with the help of offline respeaking, timecoding the output and correcting it as needed, using the Sanelius respeaking software (two tasks);
- post-editing automatically generated subtitles using Google ASR and timecoding by Flow (two tasks); and
- post-editing of automatically generated subtitles using Lingsoft ASR and timecoding by Flow (one task).

Only one clip with Lingsoft ASR was used due to delays in the production of the data and the limited availability of the participants. To account for possible differences in the difficulty of the clips, the clips were rotated between tasks. The task order was also rotated to minimise the facilitation effect.

The subtitling tasks were carried out using the subtitlers' preferred software (Wincaps Q4), and process data were collected with Inputlog (Leijten & Van Waes, 2013). To replicate the subtitlers' normal working environment, an external monitor and keyboard were used. The participants were asked to produce subtitles of sufficient quality to be broadcast, and encouraged to use all the resources they normally would (dictionaries, search engines etc.). No explicit time limit was given, and the participants were instructed to work at their own pace but to not spend excessive time on perfecting any individual subtitle or on researching information. Screen recordings were captured using software provided with Windows 10 and used to support and verify the identification of different types of activity analysed in the keylogging data.

Background information (age, length of work experience, and previous experience with ASR applications) was collected from the participants with a pre-task questionnaire. After each task, they filled out a post-task questionnaire. The post-task questionnaires were modelled after the User Experience Questionnaire (UEQ), which has been designed to elicit users' impressions, feelings and attitudes towards interactive software products (Laugwitz et al., 2008). The UEQ consists of 7-point scalar evaluations of contrastive adjective pairs (e.g. practical - impractical) describing the experience of using a product. Because our evaluation focused on the workflow and process rather than a specific software, the questionnaire was adapted by omitting adjectives related to the attractiveness or usability of the software interface. The questionnaires were provided to the participants in Finnish, with translations of the UEQ adjective pairs done by the authors.[8] Separate scalar evaluations were added for the timecoding and segmentation, and open questions were included to elicit comments on the post-editing process and the overall quality of the speech recognition.

After the completion of all tasks, a semi-structured interview was also conducted to collect more detailed feedback about problems in the workflow and the participants' views on potential improvements. The questions addressed their overall impression of the tasks, positive and negative observations, features of the ASR output which impacted the post-editing, and differences between the ASR outputs. They were also asked about their own subtitling process, how the output affected this process, whether they would use ASR in their work, and how ASR should be improved.

[8] Finnish version of the original UEQ adjective pairs was not available at the time of the first evaluation.

3.4. Proof-of-concept period in 2020

Following the first experiments, a proof-of-concept period was organised in the summer 2020. During this period the participants used the Lingsoft ASR and the automatic subtitle generation and timecoding of Flow in their daily work. The participants used these tools to subtitle whatever programmes they were working on, not restricted by genre or length. No process data were collected during the proof-of-concept period, other than self-reported approximate task times. Rather, the focus was on user experience, which the participants reported by filling a UEQ-based questionnaire (see 3.3) after each task. The open questions were reformulated to explicitly ask for comments about recurring errors in the ASR and the automatic timecoding.

3.5. Data collection for experiments in 2020

The second round of experiments was conducted in August and September of 2020, following the proof-of-concept period. Materials for these final experiments were narrowed down to a single genre, EU election debates, because the first round of experiments identified this genre as a more likely scenario for the use of ASR. As in the previous experiment round, three 3-minute clips were selected from a longer programme in such a way that they formed coherent, self-contained segments.

The participants completed three tasks: subtitling from scratch, as they would in their normal working environment, including timecoding (one task), and post-editing of automatically generated subtitles using Lingsoft ASR and timecoding by Flow (two tasks). This round used only the Lingsoft ASR, whose results had been better in the previous round. To account for possible differences in the difficulty of the clips, the clips were rotated between tasks and the task order was rotated to minimise the facilitation effect.

The subtitling tasks were conducted remotely and unsupervised due to the Covid-19 pandemic. The participants used their preferred software (Wincaps Q4) on their own work laptops, in their normal remote working environment. Due to the remote setup, no keylogging data or screen recordings could be collected; instead, the participants were asked to self-report task times to the nearest minute. The participants received detailed instructions and their tasks via email and Google Drive. Same instructions regarding subtitle quality and resource use were provided as in the first round and no explicit time limit was given. Because of the limitations caused by the remote setting, the data collected in 2020 only contribute to the qualitative part of the study focusing on the user experience.

After each completed task, the participants filled out a post-task questionnaire similar to the one used in the 2019 experiments (see 3.3), with minor adjustments made to account for the experiments being conducted remotely. In this experiment round, the questionnaire included open questions about the ASR quality, the quality of the automatic segmentation and timecoding, what the video to be subtitled was like, and whether the quality of the ASR output differed from the 2019 experiments.

After the completion of all tasks, a semi-structured interview was also conducted via video call to collect more detailed feedback about problems in the workflow and the participants' views on potential improvements. The questions addressed their overall impression of the experiment, whether automatic subtitling worked better with some tasks than others, and whether there was difference compared to the previous round. They were also asked whether they would use ASR in their work, and how ASR should be improved.

4. Results

In this section, we first examine the potential effect of automation on productivity based on the process data (section 4.1). The subtitlers' perspective is discussed based on the user experience questionnaire (section 4.2) and semi-structured interviews (section 4.3).

4.1. Process analysis: task times, keystrokes and edit distance

The effect of automatically generated subtitles on productivity was assessed in the 2019 experiments using effort indicators identified in the keylogging data. Effort indicators (task time and number of keystrokes) were compared between subtitling from scratch, subtitling with respeaking, and subtitling with different ASR outputs. For the purposes of this analysis, the Inputlog software analysis functions were used to focus only on the task time and keystrokes logged in the subtitling software, excluding other activity like online searches.

The first measure compared was the task time in each subtitling condition. Table 1 shows each participant's task times in minutes logged for subtitling from scratch, subtitling with respeaking, subtitling with Google ASR output, and subtitling with Lingsoft ASR output. Comparison of the task times shows that, on average, the participants were in fact fastest when subtitling from scratch, followed by subtitling with respeaking. Comparing the election debate content, post-editing the Lingsoft ASR content appeared to be faster than using the Google ASR, although task times varied for different clips and different participants. Subtitling the youth-oriented programme was faster, overall, than subtitling the election debate content. Differences between the participants were seen in all the task types, with participant D being consistently slower than the other three. Participant C is the only one whose average task time when post-editing ASR was lower than when subtitling from scratch.

Table 1.

Participants' Task Times (in minutes) in the 2019 Experiment

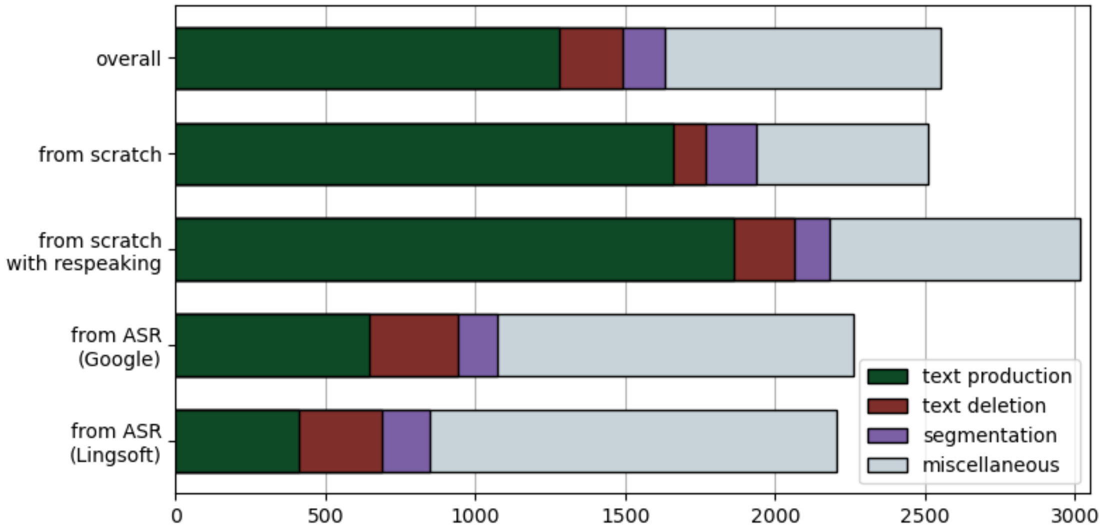
Task	Genre	Participant A	Participant B	Participant C	Participant D	Average
from scratch	debate	19:16	15:41	17:08	24:45	19:12
from scratch	youth	14:07	13:33	11:32	19:02	14:33
respeaking	debate	14:57	20:00	21:21	37:57	23:34
respeaking	youth	15:37	16:09	17:29	24:22	18:24
Google ASR	debate	20:57	28:07	18:20	31:02	24:37
Google ASR	youth	17:58	17:33	14:51	31:13	20:24
Lingsoft ASR	debate	13:40		11:37	40:13	21:50

To compare technical effort, we analysed the number of keystrokes logged in the subtitling software. Based on the process logs and Wincaps Q4 documentation, keystrokes were grouped into categories related to specific operations during subtitling. Text producing keystrokes include all alphanumeric keys that produce characters. The respeaking software is a keyboard emulator, and Inputlog cannot distinguish which keystrokes originate from the emulator and which from the physical keyboard. For this reason, the keystrokes logged by Inputlog during respeaking include the characters entered through dictation as well as those typed on the keyboard. Text deleting keystrokes include actions that remove characters. Function keys related to manipulating the subtitle blocks (creating, deleting, splitting or merging subtitles or adjusting timestamps) were grouped separately as segmentation keystrokes. Finally, keystrokes related to other actions like navigation and controlling the video were grouped as miscellaneous.

Figure 1 shows the average number of keystrokes in total and categorised by type for the four task types. The number of keystrokes was lower post-editing than for subtitling from scratch, particularly in terms of text production. Conversely, post-editing the ASR involved more text deletion to correct the output, as well as miscellaneous keystrokes, for example, arrow keys used for navigation. Comparing the two different ASR systems, post-editing the Lingsoft ASR involved slightly fewer keystrokes, with the difference again most visible in text production. The number of keystrokes is highest for respeaking, however, the technical effort is not directly comparable with the other settings since Inputlog also logs the dictated characters.

Figure 1.

Average Number of Keystrokes Grouped by Task Type and Type of Keystrokes



The number of keystrokes again varied for different clips and different participants. The slowest participant (D) used the most keystrokes overall, and also used more keystrokes when post-editing ASR output than when subtitling from scratch. The other three participants used fewer keystrokes when post-editing, although only for participant C the task times were also shorter.

Technical post-editing effort can also be observed in the number of changes made, which was assessed using two edit distance metrics. The Word Error Rate (WER), derived from Levenshtein edit distance, operates on the level of words (defined as strings of characters separated by whitespace or punctuation), and calculates the number of deletions, substitutions and insertions between the ASR output and the post-edited version, while the Letter Error Rate (LER) calculates the number of changes on the character level (see e.g. Enarvi, 2018). For calculating edit distances, the SRT files were converted into plain text, automatically joining sentences that continue from one subtitle to the next.

Table 2 shows the average WER and LER scores. For 2019, results are shown separately for the Google ASR and Lingsoft ASR as well as the election debate and youth programme clips. Edit distances for 2020 (only Lingsoft ASR and debate content) are also provided for comparison. The scores are shown as percentages where 0% indicates no changes and 100% indicates complete rewriting.

The average edit distances show that the participants changed the automatic subtitles considerably. The lowest scores are seen in the 2019 experiments for the Lingsoft output (election debate content) where less than 50% of words were changed. The Google ASR outputs were edited more, particularly in the youth programme genre. LER scores are lower in each case, suggesting that many changes involve word forms rather than replacing words. Interestingly, edit distances are higher in the 2020 experiments carried out after the participants had more experience with ASR post-editing.

Table 2.

Average Edit Distance by ASR Output Type

ASR output	Genre	Average WER (%)	Average LER (%)
2019 Google ASR	Debate	58.08	35.35
2019 Google ASR	Youth	87.44	45.21
2019 Lingsoft ASR	Debate	48.54	25.95
2020 Lingsoft ASR	Debate	81.08	50.49

In this post-editing situation, not all changes captured by the edit distances represent ASR errors. Creating the final subtitles involves also condensing to fit the space and reading time restrictions. The need for condensation was also seen when comparing the segmentation and timing of the automatically generated vs. post-edited versions. None of the automatically generated subtitle blocks were in fact accepted by the participants: 12% of subtitle blocks in the automatically generated SRT files matched either the start or the end time of a block in the corresponding post-edited file, while 88% had no match. A general tendency was to reduce the number of subtitle blocks and line breaks.

4.2. Post-task questionnaires

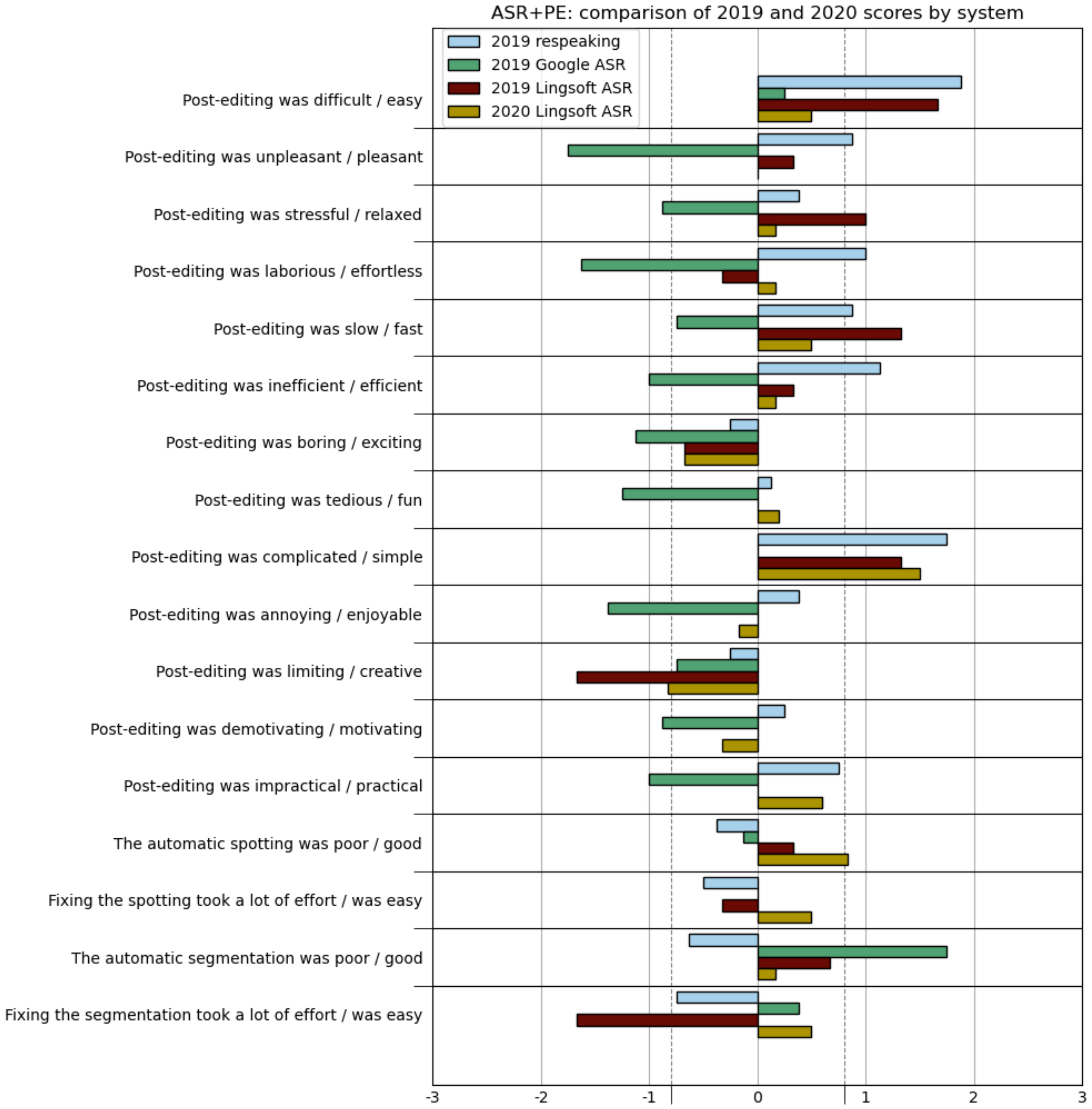
The participants' subjective evaluations about the use of automatic speech recognition and the post-editing experience were collected after each task in both rounds of the experiments and during the proof-of-concept period (see 3.3). Evaluations of respeaking in the 2019 round were also collected for comparison. We first present the questionnaire results and comments of the two experiment rounds and then feedback collected during the proof-of-concept period.

4.2.1. Results from the evaluation experiments in 2019 and 2020

Figure 2 shows the questionnaire scores for each adjective pair averaged over all participants and all clips in the post-editing and respeaking tasks of both rounds of experiments. In the figure, negative assessments (e.g. "Post-editing was laborious") are represented by bars extending toward the left side of the graph (below 0) while positive assessments (e.g. "Post-editing was effortless") are represented by bars extending to the right (above 0). According to the UEQ documentation, average scores between -0.8 and +0.8 (shown as dotted lines) are considered neutral assessments, while average values crossing these thresholds represent negative or positive assessments.

Figure 2.

Average User Experience Scores in the 2019 and 2020 evaluation experiments[9]



On average, participants scored for the Lingsoft output better than Google. Only the adjective pair "limiting/creative" is assessed more negatively for Lingsoft. Especially when post-editing the Lingsoft

[9] The participants received a questionnaire with Finnish adjective pairs. The figure shows the back-translations into English, which differ from the original UEQ wordings by Laugwitz et al. (2008).

ASR output, participants characterised the experience as relatively easy, fast, relaxed, simple and somewhat efficient, but also considered it boring and limiting. The automatic timecoding and segmentation were considered relatively poor, but easy to correct. Some differences can be seen between the individual participants' assessments: A tended to be slightly negative, B somewhat neutral, C quite positive, and D (first round of experiments only) strongly negative. Overall, participants appeared to prefer respeaking over post-editing in the first round.

In the first-round, comments provided as responses to the open questions were mainly negative, describing the post-editing process as laborious, sluggish, tiresome and unpleasant. Some comments were more mixed, for example, participants stated that post-editing was easier in the sense that it involved less typing but making the subtitles from scratch would be more straightforward. One participant characterised the process as pleasant and easy, but mechanical. Another participant observed that although much editing was needed, it was not necessarily because of poor ASR quality but because of reading speed requirements. Only participant C made explicitly positive comments, calling the post-editing process pleasant and easy and the automatic timecoding helpful.

First-round comments about the Google ASR output varied from "lousy" to "as expected" to "surprisingly good", although most comments were negative. The Lingsoft ASR output quality was described as "quite good", "relatively good", and even "very good" in both rounds of experiments. Common errors mentioned by participants involved misrecognised proper names or other words, wrong case endings, incorrect capitalisation and compound words written separately. In the second round, participants A and C found the quality similar to the first round, noting that ASR quality appeared high in both rounds. Participant B considered the output quality better in one clip.

Especially in the tasks where participants used the Lingsoft ASR, negative comments focused more on timecoding and segmentation produced by Flow than the ASR output itself. Participant D commented that the quality of the ASR output was so good that post-editing would have been faster than making the subtitles from scratch if the timecoding had been better. The participants also commented on segmentation problems like sentences being split incorrectly into subtitle blocks and lines. Although participants also mentioned having to adjust the timing of the subtitles, in the UEQ scores the automatic timecoding was mostly assessed to be good, and the errors easy to correct.

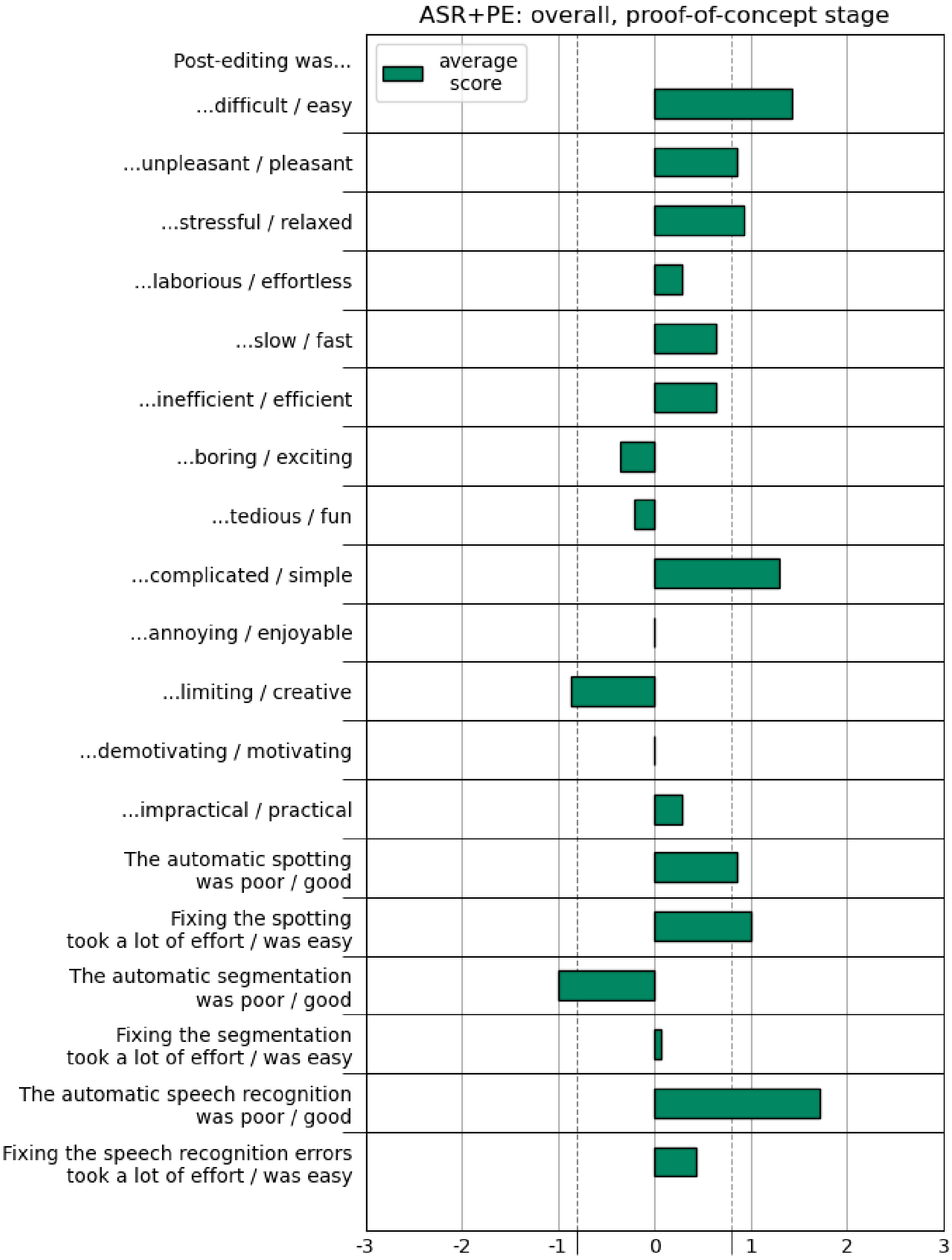
Some of the comments made by the participants addressed observations about the video clips themselves and how features of the clips affected the post-editing experience. The most common issue mentioned was the speed and style of speech: if the clip contained fast unscripted speech and rapid speaker changes, the participants needed to condense the ASR output considerably and adjust the timecoding, even if there were no ASR errors. The reading speed of the automatically generated subtitles was routinely 13-15 cps, at times up to 20 cps, while the recommended reading speed is 10-12 cps. Some specific speakers in the clips were also named as particularly difficult to subtitle because their speaking style was fast or unclear. The participants also noted that seeing the ASR output affected their processes. They sometimes struggled to find a more condensed version of what was being said and the output also steered the register toward a colloquial style.

4.2.2. Results from the proof-of-concept period

Figure 3 shows the average questionnaire scores for the proof-of-concept period. Assessments are similar to the two experiment rounds, tending toward positive. Clearly negative scores were given only for the adjective pair “limiting/creative” and for the quality of the automatic segmentation. The participants considered post-editing somewhat pleasant and relaxed, and relatively easy and simple, and their evaluation of the ASR quality was very positive. The automatic timecoding (“spotting” in the graphic) was deemed both of good quality and easy to fix. Of the three participants, A again tended toward negative, B was relatively neutral and C was the most positive. However, comparing the scores from the proof-of-concept period is complicated as the participants worked with different programme types and genres. Based on their comments, the type of the programme impacted the post-editing experience significantly.

Figure 3.

Average User Experience Scores for ASR Post-Editing during Proof-of-Concept Period



The participants' open comments about post-editing were generally mixed or neutral. ASR output was again considered helpful because it reduced the need for typing, but at the same time correcting recognition errors and editing out unnecessary words was deemed tiring and annoying. On the other hand, some positive comments noted that post-editing was fast and easy. When specifically asked about ASR errors, the participants mentioned incorrect vowel or consonant length, misrecognised proper names or non-Finnish words, compound words, problems in the presentation of numbers, incorrect capitalisation, words being incorrectly replaced by punctuation marks, and omitted words. However, many comments noted that the ASR output was overall good, and had no recurring errors.

The participants again criticised the segmentation generated by Flow, elaborating that sentences were split into multiple blocks incorrectly in a way that did not follow linguistic divisions. Other problems included missing full stops at the end of sentences, and diarisation errors (i.e. errors in separating the speakers). Incorrect segmentation also led to the need to edit the in and out times, although otherwise the participants found the timecoding overall good.

Most comments about the post-editing process related more to the programme and the speakers than the automatically generated subtitles. The participants noted that the ASR output was good when the programme had only one speaker and slow, scripted speech, but struggled with multiple speakers and fast, unscripted dialogue. The ASR output also frequently included hesitations and repetitions that could have been left out - one participant called ASR "too precise" in this sense. In addition to fast speech and unclear articulation, which were mentioned also in the experiment rounds, the participants brought up background noise as a feature that caused problems for the ASR. Like in the experiment rounds, the participants mentioned feeling limited by the wording of the output, and worried that small mistakes such as one letter differences in case endings might slip through.

4.3. Interviews

After each round of experiments, semi-structured interviews were conducted with each participant, in person for the 2019 round, and over video call for the 2020 experiments. The interview transcripts were analysed for positive and negative statements, issues raised by the participants, and suggestions for future improvements. Statements were categorised as positive if they contained favorable characterisations of the automatically generated subtitles, process of post-editing or other features of the user experience (e.g. "high quality", "easy to correct") and negative if they contained negative characterisations (e.g. "many errors", "annoying to correct"). Statements containing both positive and negative expressions (e.g. "many errors but easy to correct") were categorised as mixed comments. The numbers of negative, positive, and mixed comments are shown in Table 3. After the first round of experiments, most of the comments were negative, but many were also positive or mixed. Most of the comments were not specific to an ASR output but referred to the post-editing experience in general. After the second round, the interviews were neutral in tone, with a slight tendency towards positive. The number of mixed and positive comments was roughly the same in both rounds, but the number of negative comments in the second round was less than a third compared to the first round. This change could be because the second experiments focused on the Lingsoft ASR output, which was found more promising in the first round.

Table 3.

Number of Negative, Positive and Mixed Comments Identified in the Interviews

Comments	Negative	Mixed	Positive
Round 1 (2019)	30	8	12
Round 2 (2020)	9	9	11

Specific issues discussed in both rounds of interviews included incorrect timecoding of the subtitles, errors in recognising proper names, errors in compound words, punctuation, recognition of similar words, and omitted words. Even though errors in proper names were mentioned, the participants noted that most proper names were in fact correctly recognised. Participants' perception of the speech recognition quality appeared similar in both rounds. When asked if they noticed a difference in the quality in the second round compared to the previous, two out of three participants stated that the quality seemed similar, and the third did not recall the previous quality well enough to comment. All three participants considered the quality of the ASR output high overall.

Many of the negative comments also pertained to the effect of post-editing on the work process: the participants found that compared to their normal processes, post-editing caused more effort, was more difficult, and led to more need to navigate around the text. Comments also characterised the ASR output as frustrating and limiting. The participants made many observations about ASR struggling with spontaneous speech, colloquial style, unclear speech, and interruptions. Participant B observed that post-editing required more mental processing than subtitling from scratch, being in some ways less straightforward. Participant C mentioned concerns about using ASR, specifically that small errors in the ASR output could be easy to miss, and that post-editing could limit creativity.

Positive comments in both rounds pertained to the effect of using the ASR output. Post-editing was described as easier and even interesting, and it was said to reduce the need for typing and affect the process positively. Specifically mentioned positive features included the system's ability to omit repetition or "unnecessary" words, the correct recognition of many proper names and compound words, and overall fewer recognition errors than expected. Participant C also noted that seeing the ASR output could have a positive effect by making it easier to start the task. Participants B and C noted that getting more experience during the proof-of-concept period made working with ASR easier. Participant A stated that the experiments were interesting and expressed an overall positive view of the technologies involved. Mixed statements generally combined these views on the effect of the post-editing process: sometimes the ASR was helpful when there was not much need for corrections, but with less suitable content frequent editing was still needed.

When asked whether they would use automatic subtitles in their everyday work, if such a tool was available, Participant A first said that they would not, if the quality was like in the experiments. After the second round, Participant A adjusted their stance, stating that they would use automatic subtitles in some tasks, but not all. In both interviews, Participant B showed significant enthusiasm about the prospect of using automatic subtitles. Participant C also stated in both interviews that they would use automatic subtitles for some tasks. Participant D, who participated only in the first round, stated that they would not use automatic subtitles in their work. One participant also noted that it would be important for it to be the subtitler's choice whether and when to use these tools, because that decision would depend on more than just the genre of the programme. When asked about future improvements, the participants mentioned the same issues as in the negative comments, as well as overall improving the accuracy of the ASR. Most comments regarding future development pertained to improving the segmentation of the text into subtitle blocks.

5. Discussion and concluding remarks

The evaluations and experiments carried out in the MeMAD project suggest some promise for post-editing automatically generated intralingual subtitles as a part of professional subtitling workflows. While the participants were critical about the quality and the post-editing process, they also saw many positives, and had a positive attitude towards the technology. As was observed in Section 4, the participants' opinion of the quality of the Lingsoft ASR, in particular, was quite favorable. On the other hand, the quality of the automatic segmentation generated by the Flow platform appeared to affect the post-editing experience perhaps even more than the ASR quality, as such, based on frequent comments and improvement suggestions regarding segmentation.

With regard to productivity, findings were inconclusive. In the first experiment round, post-editing automatically generated subtitles was observed to be slower on average than respeaking of subtitling from scratch, even though the technical effort was reduced. This observation is similar to findings by Matamala et al. (2017) and Tardel (2020), while Álvarez et al. (2016b) report that all but one participant were faster with post-editing. As was discussed by the participants in our study, further practice with ASR post-editing could also lead to increased productivity (see also Álvarez et al., 2016b, p. 10844). As Tardel (2020, pp. 89-90) also points out, one explanation for reduced technical effort not leading to shorter task times is that the time needed does not consist of typing alone: when working with the ASR output, the subtitler needs to split attention between the video and transcript and to read the output before determining whether to use it or discard it. When subtitling from scratch, the subtitler may for example make decisions about omissions without needing to type them first, whereas editing the same parts out of an ASR transcript takes active effort. Deleting and rewriting passages of the transcript was observed by both Matamala et al. (2017) and Tardel (2020, p. 92), and is suggested by the number of changed words in our experiments. However, as was observed both in the comparison of the subtitle files and in the participants' comments, the number of changes is not exclusively related to ASR quality. Rather, it reflects the effort needed to edit the automatically generated subtitles so that they adhere to the guidelines adopted by the broadcaster. As can be seen in the participants' comments, a significant part of this effort was caused by segmentation issues in the automatic subtitles generated with the Flow tool. Due to these factors, metrics like number of keystrokes or number of words changed may not capture the full effort involved in using ASR for intralingual subtitling, and future research should address also the cognitive aspects of effort. Considering the participants' concern about the impact of the post-editing workflow on the quality of the subtitles, more research is also needed on that subject, with metrics that account for the necessity of editing.

Variation of effort indicators was evident between the participants, which has also been observed by Álvarez et al. (2016b), Matamala et al. (2017) and Tardel (2020). Further studies would be necessary to establish the extent to which differences in productivity are due to individual variation. The results of our study are also limited by the number of participants and the amount of data. While the number of participants is small in terms of overall generalisations, they represent a significant percentage of their own reference group. At the time of the experiments, Yle had ten full-time in-house intralingual subtitlers, so the participants represent 40% of that group in the first experiment round and 30% during the proof-of-concept period and the second experiment round.

The importance of intralingual subtitling quality is reflected in the Finnish law and guidelines to which the national broadcaster Yle is committed. Even as speech recognition improves, automatically generated subtitles cannot fulfil the guidelines as they are not able to adhere to the requirements for reading speed. Post-editing is therefore necessary to produce intralingual subtitles meeting the guidelines. Our experiments and proof-of-concept period showed some promise for this practice as all three final-round participants said that they would use automatically generated subtitles in at least some work assignments, if such a tool was available to them. However, the type of programme played a clear role in how the participants experienced post-editing of the automatic subtitles. Overall, feedback from the participants suggests that automatically generated subtitles offered most benefit when subtitling programmes with relatively slow speech and one or two speakers, but were less useful for programmes with colloquial style, fast speech and rapid speaker changes. One outcome of the project was a three-month pilot in 2021, during which all in-house subtitlers at Yle used a similar tool by Lingsoft Speech Services in their work[10]. These early steps towards implementation suggest that automation has potential as a tool for the intralingual subtitler, although further improvements in output quality and processes are needed.

Acknowledgements

The authors wish to thank the subtitlers for their invaluable participation and Umut Sulubacak for creating the graphs.

[10] A summary of the pilot can be found on Yle Sandbox YouTube channel: https://youtu.be/Suoas4r2xDI

References

Álvarez, A., Del Pozo, A., & Arruti, A. (2010). APyCA: Towards the Automatic Subtitling of Television Content in Spanish. *Proceedings of the International Multiconference on Computer Science and Information Technology*, 567–574.

Álvarez, A., Balenciaga, M., Del Pozo, A., Arzelus, H., Matamala, A., & Martínez-Hinarejos, C. D. (2016a). Impact of automatic segmentation on the quality, productivity and self-reported post-editing effort of intralingual subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 3049–3053.

Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C., & Del Pozo, A. (2016b). Automating live and batch subtitling of multimedia contents for several European languages. *Multimedia Tools and Applications, 75*(18), 10823–10853. https://doi.org/10.1007/s11042-015-2794-z

Braeckman, K., Debacq, S., Oorts, N., Van Lancker, W., Van Muylem, S., & Van Rijsselbergen, D. (2021). *D6.8 MeMAD prototype, final version*. MeMAD Project.

Doukhan, D., Francis, D., Harrando, I., Huet, B., Kaseva, T., Kurimo, M., Laaksonen, J., Lindh-Knuutila, T., Lisena, P., Pehlivan Tort, S., Reboud, A., Rouhe, A., Troncy, R., & Virkkunen, A. (2020). *D2.2 Implementations of methods adapted to enhanced human inputs*. MeMAD project. https://doi.org/10.5281/zenodo.4964299

Enarvi, S. (2018). *Modeling conversational Finnish for automatic speech recognition*. [Doctoral dissertation, Aalto University, Espoo, Finland] Aaltodoc. https://aaltodoc.aalto.fi/handle/123456789/30638

Gottlieb, H. (1997). *Subtitles, Translation & Idioms*. [Doctoral dissertation, Center for translation studies and lexicography, University of Copenhagen, Denmark.]

Kiuru, S., Koivukoski, K., & Rantanen, K. (2020). *Tekstitysvelvollisuuden toteutuminen kotimaisissa televisio-ohjelmissa tekstityksen käyttäjien näkökulmasta [How subtitling obligations come true in Finnish television programmes from the readers' point of view]* [Master's thesis, Diaconia University of Applied Sciences, Finland]. Theseus. https://www.theseus.fi/bitstream/handle/10024/336372/Kiuru_Koivukoski_Rantanen.pdf

Kóbor-Laitinen, Z. (2021). Huonokuuloisten katsojien näkemyksiä eduskunnan kyselytuntien automaattisesta ohjelmatekstityksestä [Views of hard-of-hearing audience on the automatic intralingual subtitling of parliamentary question time]. *MikaEL Electronic Journal of the KäTu Symposium on Translation and Interpreting Studies, 14*, 139–155.

Koponen, M., Tuominen, T., Hirvonen, M., Vitikainen, K., & Tiittula, L. (2020). User perspectives on developing technology-assisted access services in public broadcasting. *Bridge: Trends and Traditions in Translation and Interpreting Studies, 1*(2), 47–67. https://www.bridge.ff.ukf.sk/index.php/bridge/article/view/8

Kuuloliitto. (2019). *Kuuloliiton kannanotto sähköisen viestinnän lainsäädännön kehittämis- ja muutostarpeista [Statement regarding the development needs of electronic communications legislation, Ministry of Transport and Communication]*. Kuuloliitto. https://www.kuuloliitto.fi/wp-content/uploads/2019/03/KHL-Kannanotto-LVM-viestintäpalvelulaki-010319.pdf

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In A. Holzinger (Ed.), *HCI and Usability for Education and Work. USAB 2008* (Vol. 5298, pp. 63–76). Springer. https://doi.org/10.1007/978-3-540-89350-9-6

Leijten, M., & Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication, 30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Liu, D. (2014). On the classification of subtitling. *Journal of Language Teaching & Research*, 5(5).

Matamala, A., Oliver, A., Álvarez, A., & Azpeitia, A. (2015). The Reception of Intralingual and Interlingual Automatic Subtitling: An Exploratory Study within the HB4ALL Project. In J. Esteves-Ferreira, J. Macan, R. Mitkov, & O.-M. Stefanov (Eds.), *Translating and the Computer 37*, pp. 12–17. ASLING. https://doi.org/10.7202/002966ar

Matamala, A., Romero-Fresco, P., & Daniluk, L. (2017). The use of respeaking for the transcription of non-fictional genres: An exploratory study. *InTRAlinea, 19*, (no pagination).

Orrego-Carmona, D., Dutka, Ł., & Szarkowska, A. (2018). Using translation process research to explore the creation of subtitles: an eye-tracking study comparing professional and trainee subtitlers. *The Journal of Specialised Translation, 30*, 150–180.

Piperidis, S., Demiros, I., Prokopidis, P., Vanroose, P., Hoethker, A., Daelemans, W., Sklavounou, E., Konstantinou, M., & Karavidas, Y. (2004). Multimodal Multilingual Resources in the Subtitling Process. *Fourth International Conference on Language Resources and Evaluation* (LREC 2004).

Romero-Fresco, P. (2011). *Subtitling Through Speech Recognition: Respeaking*. St. Jerome.

Tardel, A. (2020). Effort in Semi-Automatized Subtitling Processes: Speech Recognition and Experience during Transcription. *Journal of Audiovisual Translation, 3*(2), 79–102. https://doi.org/10.47476/jat.v3i2.2020.131

Tiittula, L., Kurimo, M., Mansikkaniemi, A., & Rainò, P. (2018). Ohjelmatekstityksen toimivuus eri kohderyhmien näkökulmasta [Functionality of intralingual subtitling from the perspective of different target audiences]. *MikaEL Electronic Journal of the KäTu Symposium on Translation and Interpreting Studies, 11*, 20–34.

Vitikainen, K. (2018). Developing Live Subtitling in Finland: Moving from manual subtitling towards respeaking. *MikaEL Electronic Journal of the KäTu Symposium on Translation and Interpreting Studies, 11*, 35–48.

Vitikainen, K., Lehikoinen, K., Holopainen, T., Ristola, T., Pöntys, M., Kauppila, J., Stenbäck, M., Korhonen, R., Metsola, K., Lehto, L., Häkkinen, T., Benigni, A., Gorschelnik, H., & Antinjuntti, K. (2020). *Ohjelmatekstitysten laatusuositukset [Quality guidelines for intralingual subtitles].* Kieliasiantuntijat. https://kieliasiantuntijat.fi/wp/wp-content/uploads/2021/01/Ohjelmatekstitysten_laatusuositukset_web-versio.pdf