Journal of Audiovisual Translation

Effort in Semi-Automatized Subtitling Processes: Speech Recognition and Experience during Transcription

Anke Tardel✉
Johannes Gutenberg University Mainz

Citation: Tardel, A. (2020). Effort in Semi-Automatized Subtitling Processes: Speech Recognition and Experience during Transcription. *Journal of Audiovisual Translation, 3*(1), 79–102.

Editor(s): A. Matamala & J. Pedersen

Received: February 07, 2020

Accepted: August 24, 2020

Published: December 18, 2020

Copyright: ©2020 Tardel. This is an open access article distributed under the terms of the Creative Commons Attribution License. This allows for unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

The presented study investigates the impact of automatic speech recognition (ASR) and assisting scripts on effort during transcription and translation processes, two main subprocesses of interlingual subtitling. Applying keylogging and eye tracking, this study takes a first look at how the integration of ASR impacts these subprocesses. 12 professional subtitlers and 13 translation students were recorded performing two intralingual transcriptions and three translation tasks to evaluate the impact on temporal, technical and cognitive effort, and split-attention. Measures include editing time, visit count and duration, insertions, and deletions. The main findings show that, in both tasks, ASR did not significantly impact task duration, but participants had fewer keystrokes, indicating less technical effort. Regarding visual attention, the existence of an ASR script did not decrease the time spent replaying the video. The study also shows that students were less efficient in their typing and made more use of the ASR script. The results are discussed in the context of the experiment and an outlook on further research is given.

Key words: speech recognition, effort, transcription, eye tracking, keylogging, subtitling processes.

✉ antardel@uni-mainz.de, https://orcid.org/0000-0001-9282-5824

1. Introduction

Audiovisual productions contribute tremendously to Europe's economy and culture. According to the European Audiovisual Observatory (Schneeberger, 2019), Germany is among the three largest audiovisual markets with 370 channels. The main asset of the audiovisual market is the content of TV production companies. While the immediate market size of audiovisual content is limited by the production language, other markets can be reached by re-purposing the content via translation. Audiovisual translation (AVT) is mostly done via dubbing, voice-over, or subtitling. Dubbed versions are complex and expensive to produce, so a Europe-wide distribution of local audiovisual programmes cannot be achieved for all productions (Raats, Evens, & Ruelens, 2016). This depends on the distribution platform, and on the practices in individual countries (Baños & Diaz-Cintas, 2017).

Over time, more TV productions have found their way into the catalogues of streaming providers. This introduces additional financing opportunities for (co-)producers and licence holders of new programmes. The distribution via video on demand platforms has become a significant factor in the (co-)financing of new programme ideas, opening new market opportunities, particularly for independent producers, beyond mere financing of broadcast stations (Raats et al., 2016). The widespread availability of digital providers results in an intensive pan-European cultural exchange through the content of various genres. Most language adaptations are implemented through subtitles or dubbing. Subtitling provides the advantage of retaining the original language. Additionally, subtitles are a proven means of accessibility for hearing impaired, but also language learners, immigrant communities, and people without access to the audio (Diaz-Cintas & Remael, 2014, p. 14).

Subtitling is generally cheaper than dubbing depending on the country and its working traditions (Media Consulting Group & Peacefulfish, 2007, p. 34). Creating large programme packages in numerous language combinations, even with subtitles only, requires innovations in production processes to adequately cope with market changes caused by digitization. The production of subtitles depends on three parameters: volume, deadlines, and price (Media Consulting Group & Peacefulfish, 2007, p. 74). Many steps of the subtitling process at German broadcasting companies are still often carried out by few individuals in relatively unautomated or unformalized environments. Conventional technical aids support individual steps, but not in the form of integrated or synergistic technical environments or platforms. Such platforms could lower costs by providing central upload of source and reference material on the producer end and assistive technology such as automatic speech recognition (ASR), neural machine translation (NMT), translation memories (TMs), glossaries, and configuration of general and client-specific style guides on the user end.

2. Empirical Research on Speech Recognition in AVT

Recently, we have seen a rise in platforms integrating upload processes, language technology and tools for quality assurance, but they have yet to be scrutinized by in-depth research (Díaz-Cintas,

2013; Cintas & Massidda, 2019). Neural networks enable the design of an overall workflow for subtitling content featuring an open architecture with Application Programming Interfaces (APIs) for ASR and NMT from different vendors. A sensible balance of human skills and computer support are both qualitatively promising and more economical, sufficiently agile, and future-proof.

ASR is the automatic process "of converting a speech signal to a sequence of [written] words, by means of an algorithm implemented as a computer program" (Anasuya & Katti, 2009, p. 181). This contrasts with manual transcription where the transcriber listens to and watches the audiovisual content while typing the dialogue without further assistance. An already common practice in live-subtitling is the use of ASR in combination with respeaking. An ASR system can be trained to be speaker-dependent, enabling it to understand the respeaker's voice even with difficult vocabulary. Speaker-independent ASR is needed when the audio of a video with multiple speakers is to be automatically transcribed without respeaking. Here, the recognition rate depends on training data, audio input quality, number of speakers, background noise, and context. In both approaches, human revision is necessary. The objective is to create ideal technical prerequisites through standardization for language processing and to achieve an optimization of the results through post-editing (PE) the ASR output similar to the process with output of machine translation (MT). For these processes, crowdsourcing and an effective integration into the workflow are viable options. If transcripts can be created cost- and time-efficiently with PE to be further processed by MT, these transcripts can assist subtitlers in their work.

As speech recognizers are learning systems, a continuous optimization can be expected, especially when edits during PE are fed back into the system. This, however, could only work with verbatim transcripts, but not with condensed subtitles. To our knowledge only one study has specifically compared manual transcription of audiovisual content with ASR and transcription via respeaking within the ALST project (Matamala, Romero-Fresco, & Daniluk, 2017). In this small-scale study, temporal and perceived effort was measured during transcription under three conditions. They found a tendency of manual transcription being the fastest and PE of ASR the slowest of the tested methods. More studies like this with different ASR systems, genres, languages, and more participants are necessary. In subtitling, ASR has been tested mainly in live-subtitling (e.g., Aliprandi et al., 2014) or automatic subtitling of video lectures (e.g., Quintas, 2017).

There are still gaps in empirical research concerning computer-assisted subtitling. It remains unclear whether a corrupt ASR transcript might still facilitate subtitling processes, and under which circumstances, or the role which the visual part of the video plays during computer-assisted transcription and translation. Some of the challenges and potential possibilities for automatizing the subtitling process were addressed by the COMPASS[1] project, discussed in the following section.

[1] See COMPASS project site https://www.compass-subtitling.com/

3. The COMPASS Project

The project COMPASS, managed by ZDF Digital and Johannes Gutenberg University of Mainz, was funded by the European Commission with the aim of optimizing multilingual subtitling processes for offline public TV programmes. By reviewing workflows in subtitling, leveraging state-of-the-art ASR and NMT, the conventional workflow for the creation of translated subtitles was transformed into a uniform process model within a platform that is automated wherever possible. The focus has been to allow technology, human-machine interaction, and machine learning approaches to optimize and continuously improve processes at crucial points: automatic ingest and material supply, transcription and translation, compliance with platform-specific standards, and quality assurance.

With growing research interest and advances in ASR and NMT and their increasing application in the captioning online courses, e.g., TraMOOC (Kordoni et al., 2016) and TransLectures (Silvestre Cerdà et al., 2012), it is time to consider implementing these technologies in TV subtitling as well. Within previous research projects such as SUMAT[2] , MUSA[3] , and eTITLE (Álvarez, Arzelus, & Etchegoyhen, 2014; Del Pozo et al., 2014; Melero, Oliver, & Badia, 2006), these approaches have not been studied with a focus on the process, but rather they have looked at comparing outputs and perceived effort.

The project's focus was on combining human and machine input to make the process of interlingual subtitling as efficient and fit for purpose as possible. Post-editing of machine translation (PEMT) is standard in the translation industry, but typically not used for offline subtitling, although the industry is adapting (Bywood , Georgakopoulou, & Etchegoyhen, 2017). PE for instance is sometimes used in intralingual subtitling, when a live-subtitled programme is corrected, resynchronized, and uploaded to an online platform. This is not common practice for all programmes and interlingual subtitles. The planned COMPASS pipeline foresees the use of ASR to extract a film transcript, followed by human PE of the ASR texts. The transcripts are then roughly automatically timecoded and manually converted into monolingual subtitles. If the programme is also to be interlingually subtitled, the PE transcripts (or subtitles) are then translated via NMT into English as relay language and the target languages. The question is whether it is more efficient to first post-edit the MT transcript to assist the subtitler, or to apply PE MT directly to subtitles. The use of template files is not new in subtitling, but there are ongoing discussions on quality and information loss (Georgakopoulou, 2019). Semi-automatic, interlingual transcripts could be a solution, and open new research avenues. Based on these considerations, this study examines transcription processes with and without the assistance of ASR.

[2] See SUMAT project site http://www.fp7-sumat-project.eu/
[3] See MUSA project site http://sifnos.ilsp.gr/musa/

4. Methodology

This study applies established methods from translation process research (TPR) to subprocesses of subtitling, i.e., verbatim transcription and translation of film dialogue as in the "written representation of audible speech" (Matamala et al., 2017, p. 2). The focus is on the impact automatic transcripts have on students' and professional subtitlers' transcription and translation behaviour. The behavioural data is triangulated with data from questionnaires and eventually quality annotations. This article will exclusively focus on measures of effort. The study design is motivated by the intended pipeline of the COMPASS tool featuring support via ASR, NMT and translation via pivot language. The triangulation of gaze data, typing activities, and product data in linear mixed models allows a detailed analysis of the processes on different levels considering participant- and text-inherent variances. Effort is measured on the three levels proposed by Krings (2001) who applied them to compare PE and translation processes. Krings distinguishes between temporal, technical, and cognitive effort. While the temporal effort is the task completion time, technical effort describes the use of mouse or keyboard, and cognitive effort describes cognitive processes such as monitoring, reading and attention distribution. For an overview of the variables see Table 3. As subtitlers, in contrast to regular translators, must translate from oral modality to written, another focal point was the difference between professional subtitlers and translation students. Based on this and the aim of the project, the hypotheses for this study are as follows:

1) Effort during transcription differs according to experience in that translation students require more effort on all three levels than professional subtitlers.
2) Automatization approaches in transcription and translation facilitate the processes in that less effort is required on all three levels both in intralingual and interlingual tasks.

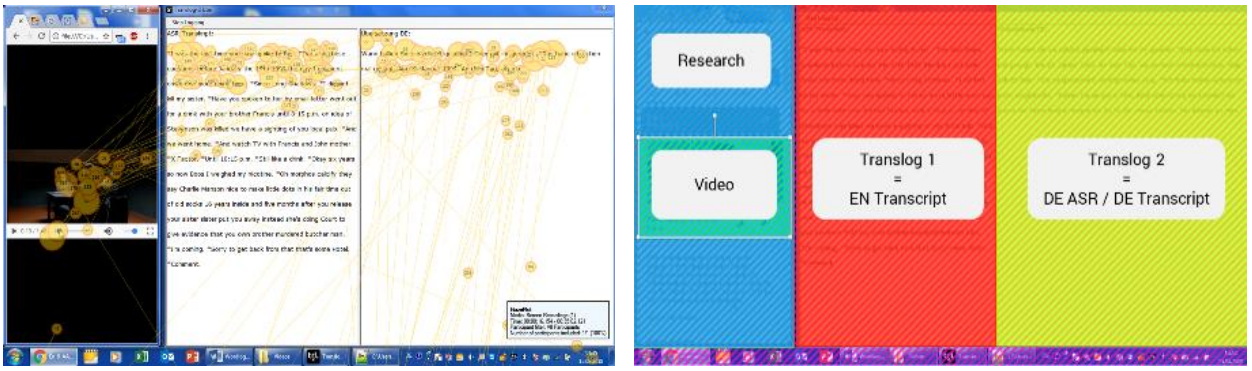
The reception of subtitles has been the subject of various eye tracking (ET) studies. The process of subtitling, however, has yet to be researched with empirical methods beyond questionnaires or case studies. Hvelplund (2017), investigating attention distribution and cognitive effort in translation for dubbing, and Orrego-Carmona, Dutka,& Szarkowska. (2018), comparing student and professional subtitlers regarding effort in interlingual subtitling, are some of the few researchers to apply keylogging and ET to interlingual AVT. Together with Beuchert (2017) they proposed the research field Subtitling Process Research, where this study fits in.

The underlying assumption of ET is the Eye-Mind-Hypothesis (Just & Carpenter, 1980). It is based on the idea that there is an immediate relationship between our focus of visual attention and the object of our cognitive attention. An eye tracker records gaze samples which are divided into fixations and saccades with a fixation filter. Fixations are cumulative gaze data points in very close proximity when the eye is assumed to be nearly still, and information is processed. The counterpart is saccades; rapid eye movements between fixations during which it is assumed that no information is processed, at least not within a complex cognitive task such as reading (Rayner, 2009). Fixation measures are analysed with an area of interest (AOI).

In this study, a Tobii TX300 remote eye tracker was used in combination with the software Tobii Studio (3.3) and a plugin for Translog-II (2.0, Carl 2012). The standard Tobii Fixation Filter was applied. Tobii Studio was used for screen recording, ET and drawing of AOIs as shown in Figure 1. These include the Google Chrome video player, and the browser window, both of which were activated only during video replay or online research. The other AOIs were drawn on the editor window(s) of Translog-II, where participants wrote the transcripts. The Translog AOIs were activated from the time participants clicked into the window until participants clicked a button to stop logging. Calibration was performed first in Tobii Studio and afterwards in Translog-II in a 5-point calibration and participants were seated at 60 cm from the eye tracker. In this analysis, only the ET data from Tobii Studio was analysed.

Figure 1

Setup on Screen with Gaze Plots (left) and AOIs in Tobii Studio (right)



All tasks were carried out in Translog-II (Carl, 2012) a screen-based keylogging tool that does not log actual keys presses, but the characters appearing/disappearing on screen, and cursor movements. It consists of a text editor with an editable target window only, or both a source (Translog 1, T1) and editable target text 9Translog 2, T2) window. Data is recorded in a format that allows post-processing via alignment of ST and TT segments and words, and the integration in the Translation Process Research Database (TPR-DB, cf. Carl et al., 2016). The type of data recorded in this study is presented in Table 3.

5. Study Design and Procedure

The study consists of three tasks in eight conditions which were subprocesses of subtitling: intralingual verbatim transcription (Intralingual), interlingual verbatim transcription (Translation), and PE via English as pivot language. The focus in this article is on the first two tasks as the analysis of the PE task is discussed in Tardel (Forthcoming). For each condition, the tasks were modified by introducing ASR or human transcripts of the videos as indicated in Table 1.

Table 1

Overview of Tasks and Conditions

Task	Language	Condition	Description	Recordings
Intralingual	DE	I	Transcription	22
Intralingual	DE	I+ASR	Transcription with ASR script	22
Translation	EN	T	Interlingual Transcription	21
Translation	EN	T+ASR	Interlingual Transcription with ASR script	23
Translation	EN	T+S	Interlingual Transcription with human script	23

Participants performed all tasks and conditions in the same order with videos alternating in a balanced pseudo-random fashion. Task 1 contained two intralingual transcription conditions and Task 2 three interlingual transcription conditions. The source language was either German or English, and the target language was German.

Prior to recording, participants were informed about the methodology, and filled out a consent form as well as a metadata questionnaire. This includes language and training background regarding subtitling and translation experience. Additionally, participants were provided with the titles of the TV series and received instructions on formatting, e.g., speaker indication with an asterisk and ellipses for incomplete utterances. Although participants were not used to this, it did not seem to impact participants in a negative way and the rules were the same for everyone.

During a copying task in Translog-II, participants became familiar with the setup, adjusted the headphones volume, and tested the video navigation. Calibrations were performed prior to every new session to ensure comparable data quality. The sessions combined took roughly three hours per participant with short breaks in between to avoid tiring effects. Subtitlers usually work on much longer tasks, which is problematic for eyetracking. Therefore, video sequences were kept comparably short. This poses a limitation to this study; nevertheless, already finding effects in shorter clips suggests that it is worth carrying out more time-consuming and data-intense studies in the future.

5.1. Sampling

Participants were sampled via convenience sampling from two groups: translation students (S; N=13) from the translation studies programme at University of Mainz, and professional subtitlers (P; N=12) working in the Berlin area. Participants were German native speakers with English as their active working language, and they all received renumeration. The bias for female participants in both datasets reflects the current market and university trends and is not expected to impact results.

The 12 female and one male translation students ranged from being in their 3rd to 6th semester (SD=4.4). Three students were in the MA translation while the rest was in their BA studies and none had substantial experience in subtitling or PE.

The nine female and three male professional subtitlers had an average of 6.7 years (Min=2 years) of professional experience in interlingual subtitling. All had either formal training in translation or in AVT and are currently working as in-house or freelance subtitlers.

5.2. Material

The videos were short scenes from two different crime series that were comparable in audio quality, length, text content and number of speakers. They made up coherent scenes taken from different episodes. None of the participants were familiar with either of the two productions: *You Are Wanted* (German, Amazon Prime, 2016) and the BBC series *River* (English, 2015). The final five scenes are presented in Table 2. The assisting transcripts were either human or automatically created (ASR) with Google Cloud Speech-To-Text. While an improved video model for audiovisual files with multiple speakers was available for the English texts, there was only a standard model for German. ASR scripts included punctuation and speaker changes. For a comparison of the ASR scripts, the transcription word error rate (WER) was computed indicating the minimum edit distance between the transcription and the reference. Two WER scores were computed to account for alignment errors in the algorithm caused by differences in punctuation and capitalization. While all edits were weighted equally, in the normalized score, all punctuations and capitalizations were removed. The scores presented in Table 2 show that the WER was lower for the English ASR. As a reference, a human transcriptionist has an average WER of 0.04 while commercial ASR average at around 0.12. The WER scores of the ASR in this study are rather poor, given that for the German ASR more than half of the script's words contained errors or were not recognized.

Table 2

Metadata on Video Excerpts Used as Source Texts

Video ID	Series	Duration (min)	Words	Words per min (wpm)	WER ASR	WER ASR (normalized)
2	*You Are Wanted*	1:48	178	99	0.79	0.69
4	*You Are Wanted*	1:48	171	95	0.7	0.54
6	*River*	1:49	188	103	0.38	0.17
8	*River*	1:48	182	101	0.4	0.15
10	*River*	1:49	189	104	0.54	0.32

5.3. Data Analysis

The statistical analysis was performed in R (R development core team 2017, version 3.6) with Linear Mixed Effect Models (LMMs) using the packages languageR (R Core Team, 2019), and lme4 (Bates et al., 2015). LmerTest (Kuznetsova et al., 2019) was used to calculate the estimate (β), standard error (SE), degrees of freedom (df), t-value as coefficient divided by its estimate (t), and significance level (p). Significant levels are highly significant (p<0.001), significant (p<0.05), or marginally significant (p<0.1). For more robust models, the dependent variables (DV) were log-transformed, to achieve close to normal distribution. Outliers larger than 2.5 standard deviations per condition were excluded. The respective datapoints are indicated in parentheses together with the other measures. The effects are visualized in plots for a better interpretation by applying the ggplot2 package (Wickham, 2016). The model fit was tested by checking the distribution of residuals. Collinearity was assessed by inspecting variance inflation factors for the predictors; all values were relatively low (<2).

In the LMMs the different DV presented in Table 3 were included. Predictors were always condition (script or ASR) and/or status (translation student, professional subtitler) to investigate the effect of support, and participant experience on effort. Random variables always included participant and item (video).

Table 3

Overview of Significant Models with Dependent Variables Describing Temporal (1), Technical (2-5), and Cognitive (6-10) Effort

Model (LMM-)	Dependent Variable (DV)	Variable Description	Measure
I1 & T1	Duration	Session duration, from first click into Translog-II to "stop logging"	min
I2	Insertion	Total number of insertions,	count
I3	Deletion	deletions, and	
	Keystrokes	total keystroke count (ins+del) per session	
I4 & T4	PDur	Total duration of continuous typing (pauses <1s)	min
I5 & T5	PNum	Total number of production units, periods of continuous typing	count
T6a	VisitDurT1	Average visit duration on T1/T2-AOI	ms
I6b & T6b	VisitDurT2		
I7 & T7	Visit CountT1	How often participants entered the	count
	Visit CountT2	T1/T2-AOI	
T8a	TrtS	Total reading time on ST/TT: Sum of fixation	min
I8b & T8b	TrtT	duration of all fixations on T1/T2-AOI	
I9 & T9	Factor Video	Video replay duration (video-AOI is active), expressed as factor of video duration	factor of video duration
I10 & T10	Relative Video	Video replay duration divided by the session duration	percentage

6. Results and Discussion

This section presents the results in various LMMs as indicated in Table 3. The results sections are subdivided into the two tasks: intralingual transcription (I) and translation (T). Within each task only significant effects of condition and status in interaction with each other or other variables are discussed.

6.1. Temporal Effort

Temporal effort describes the session completion time from the time participants first clicked into the Translog-II editor to clicking "stop logging". Across participants and conditions, participants took an average of 14 minutes per session. The shortest session was an intralingual transcription session

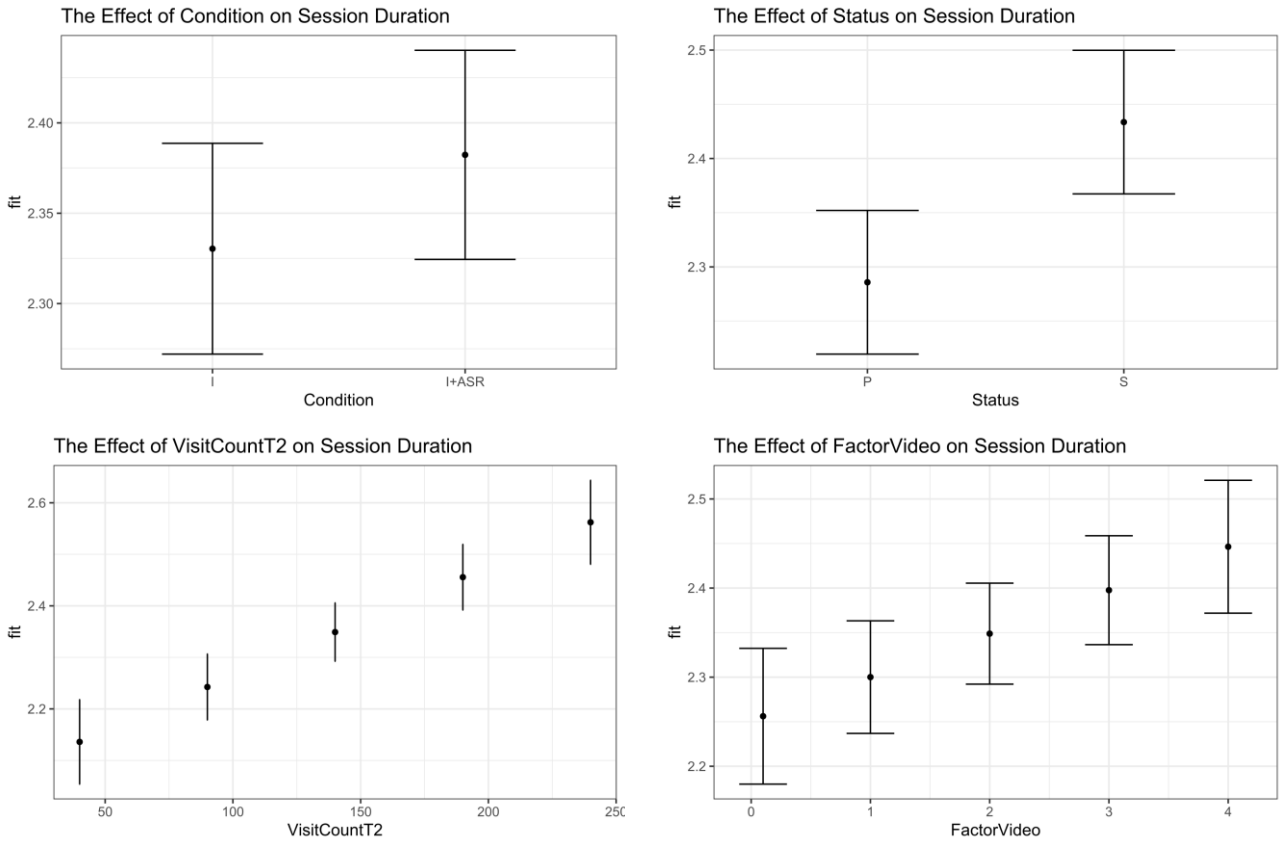
without ASR (6:47 minutes), and the longest recording was that of a translation from scratch task (almost 42 minutes).

6.1.1. Intralingual Transcription

The average temporal effort within intralingual transcription was 10:49 minutes (Min=06:47; Max=16:23; SD=02:36). The first model LMM-I1, visualised in Figure 2, contains session duration as DV, and status, condition, visit count on T2 and factor video replay as predictors. There is a significant positive effect on session duration for status student (44 datapoints: β=0.1, SE=0.07, df=24, t=2.2, p<0.05) and visit count on T2 (β=0.002, SE=0.0006, df=37, t=3.6, p<0.0001). Students, thus, took longer than professional subtitlers, while switching attention away from the TT more often also resulted in longer completion times. Condition had only a marginal positive effect (β=0.05, SE=0.02, df=26, t=2, p<0.06) just like the factor of video replay (β=0.04, SE=0.02, df=24, t=1.9, p<0.07). Thus, the task was possibly slowed down by longer video replay durations and having an ASR transcript.

Figure 2

Effect of Several Predictors on Session Duration in LMM-I1 in Intralingual Transcription



In both conditions, the students were slower than the professionals, which seems plausible as professional subtitlers are more used to decoding audiovisual content; translation students mainly work with written texts. The marginal slowing effect of ASR goes against the hypothesis that it would

support the process. The conclusion that participants switching attention between the TT window and the video or browser more often slows down their process seems plausible. It suggests that it is more efficient to continuously focus on the video or text and memorize longer stretches. When faced with an ASR script, participants had to check the script before deciding to work with it or reject it. The longer and more often participants replayed the video, the slower they were. In the replays, it was observed that participants often rejected the ASR and worked from scratch, as they were not specifically instructed to work with the ASR. When considering the WER rate of the ASR scripts, this behaviour is not surprising. With more than 50% of words containing errors in the ASR, participants decided it to be more efficient to work from scratch. More research will have to be carried out to examine the role of ASR in different languages and the quality threshold. Also, it might be interesting to look at the strategies applied. Further, respeaking might have improved the WER in the ASR, but respeaking requires yet another skillset, and training of the ASR. The idea was to include the ASR as an automatised step.

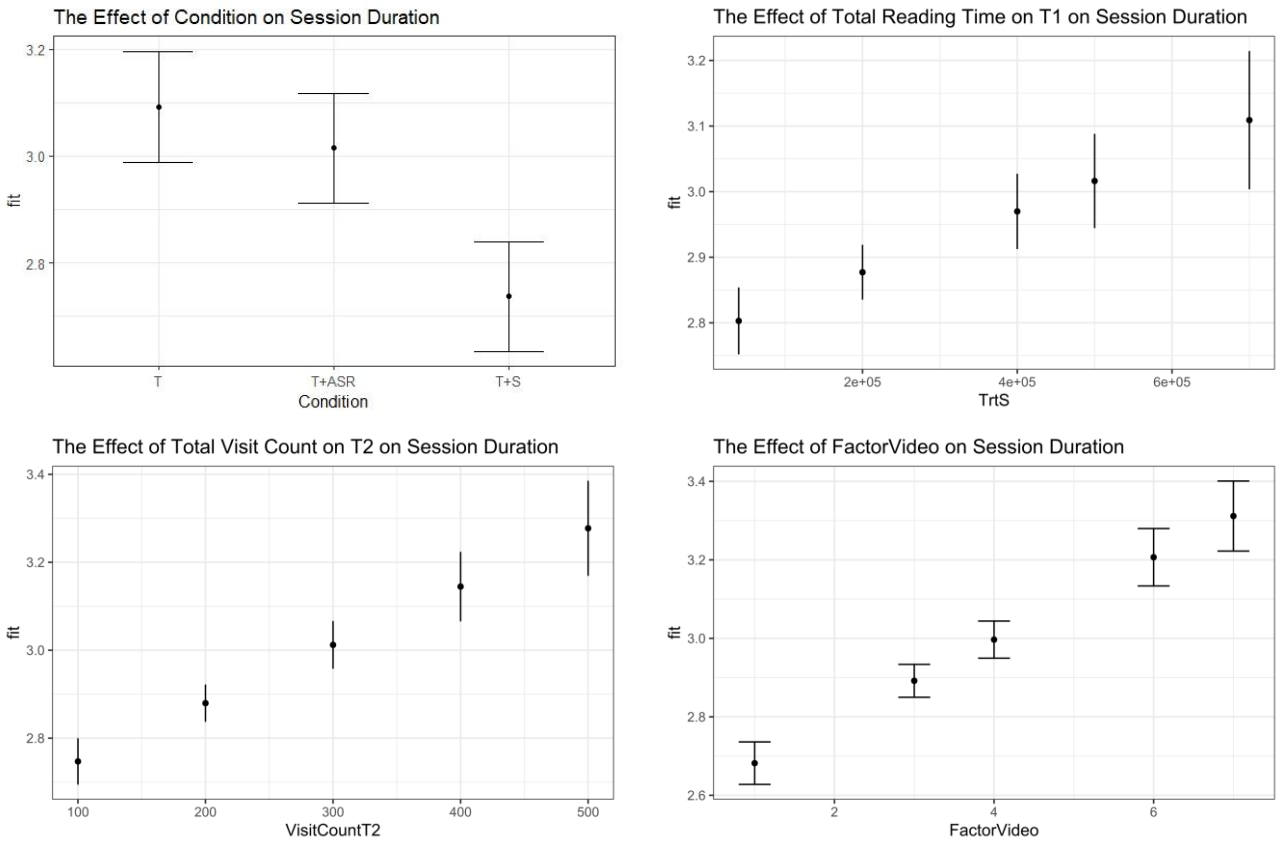
6.1.2. Translation

The average temporal effort for translation was 19:59 minutes (Min=09:09; Max=41:56; SD=03:24). The first model LMM-T1 shows a significant negative effect for working with an English script on session duration (67 datapoints: β=-0.2, SE=0.05, df=46, t=-3.3, p<0.001). As can be seen in the first plot in Figure 3, working with the ASR showed no significant effect; only a correct English transcript sped up the process. The effect of total reading time on the assisting script (67 datapoints: β=0.06, SE=0.02, df=38, t=2.4, p<0.01), the factor video replay (67 datapoints: β=0.1, SE=0.02, df=38, t=5.5, p<0.001) and the visit count on the TT (67 datapoints: β=0.001, SE<0.01, df=29, t=4, p<0.001) were all significant and positive on session duration. The longer participants read the assisting script, the slower they were, similar to longer video replay and more switching away from the TT.

Contrary to intralingual transcription, status did not have a significant effect and did not interact with the condition. Here, it seems that the temporal effort was similar for both groups. The significant negative effect only for translation with correct script, suggests that the ASR was not helpful, or hindering compared to translation from scratch. As visualised in Figure 3., the hypothesis that having an ASR script would lower temporal effort in translation could not be confirmed. Although the WER scores for the English ASR was better, there were still too many errors for it to be significantly helpful. Most errors (61-78 per video) were made up of substitutions in that the ASR recognized the wrong words, while deletions, i.e., not-recognized words, ranged from 10-22. Participants were not able to correct the ASR output for a correct written representation of the video audio. It is possible that the ASR would have been more helpful if this option were available; an aspect worth investigating in another setup.

Figure 3

Effects of Several Predictors on Session Duration in LMM-T1 in Translation



6.2. Technical Effort

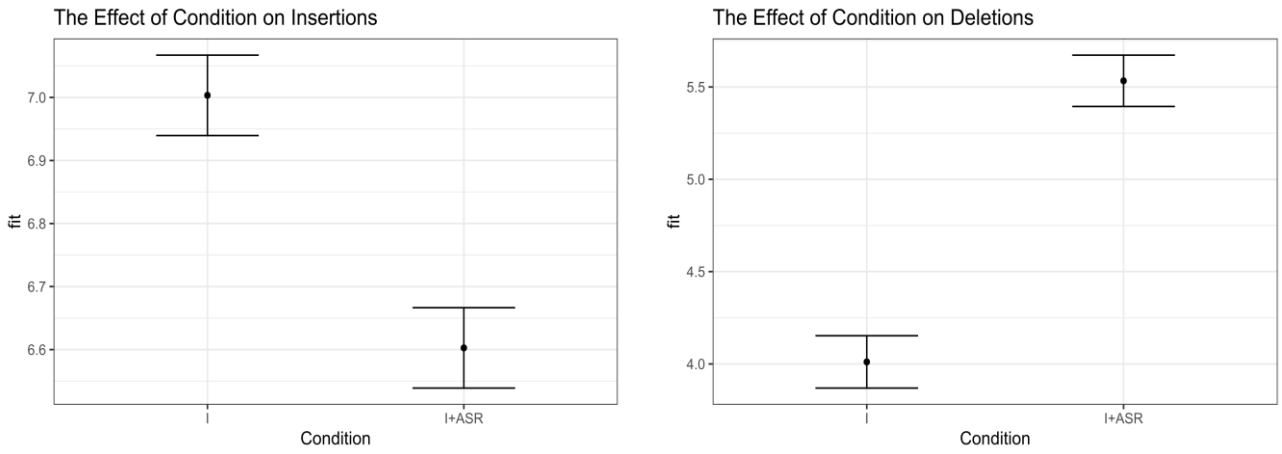
The technical effort is the amount of interaction in the form of direct keystroke measures (insertions, deletions, and keystrokes) as well as units of continuous typing and overall typing duration. Deletions and insertions covariate but describe separate forms of interaction, as only deletions directly correspond to revision. Continuous typing is defined as keystrokes that occur within one second of time. Whenever there is a writing pause for longer than one second, a new production unit is counted when the participant continues typing. These production units (Carl, Schaeffer, et al., 2016) can be an indicator of less cognitive effort as the participant does not have to think about the next keystroke. The measure of one second was applied to all participants, aware that this might be problematic as participants with slower typing speeds might be penalized. However, adding participant as random effect accounts for those differences. In general, the more of these production units are counted in a session (PNum), the more often typing was interrupted – for video consultation, online research, reading of reference script, or pausing to think about the next word to type. Another measure linked to production units is production time (PDur) which indicates the overall time spent with typing.

6.2.1. Intralingual Transcription

In LMM-I2 the DV is insertions and in LMM-I3 deletions. In both models, as can be observed in Figure 4, significant effects were found for condition, but not status. Technical effort, thus, does not differ significantly from that of professional subtitlers. In LMM-I2, the effect on transcription with ASR is negative; when working with the ASR, fewer insertions were performed (44 datapoints: β=-0.4, SE=0.04, df=22, t=-9.9, p<0.001) with only a marginal effect on TT length. In LMM-I3 the effect is significant, quite large, and positive for deletions (43 datapoints: β=1.5, SE=0.15, df=18, t=-10.53, p<0.001) in that more deletions were performed. This can be explained with participants deleting parts or the entire ASR-script and not just local revision of passages in the text. While these strategies are not considered in this analysis, it might be a promising avenue for further analysis and explain the large positive effect for deletions. The negative effect for insertions in the ASR condition (see Figure 4) suggests that the ASR condition decreases technical effort. Considering again the WER analysis, there were 38-52 correctly recognized words and if neglecting punctuation and capitalization even 56-80 words. This is less than half of the words but could explain the decreased technical effort. With typing data combined (Keystrokes) condition ASR had a significant negative effect (43 data points: β=-0.1, SE=0.05, df=22, t=-2.1, p<0.05) in that overall technical effort was lower.

Figure 4

Effects for LMM-I2 and LMM-I3 in Intralingual Transcription



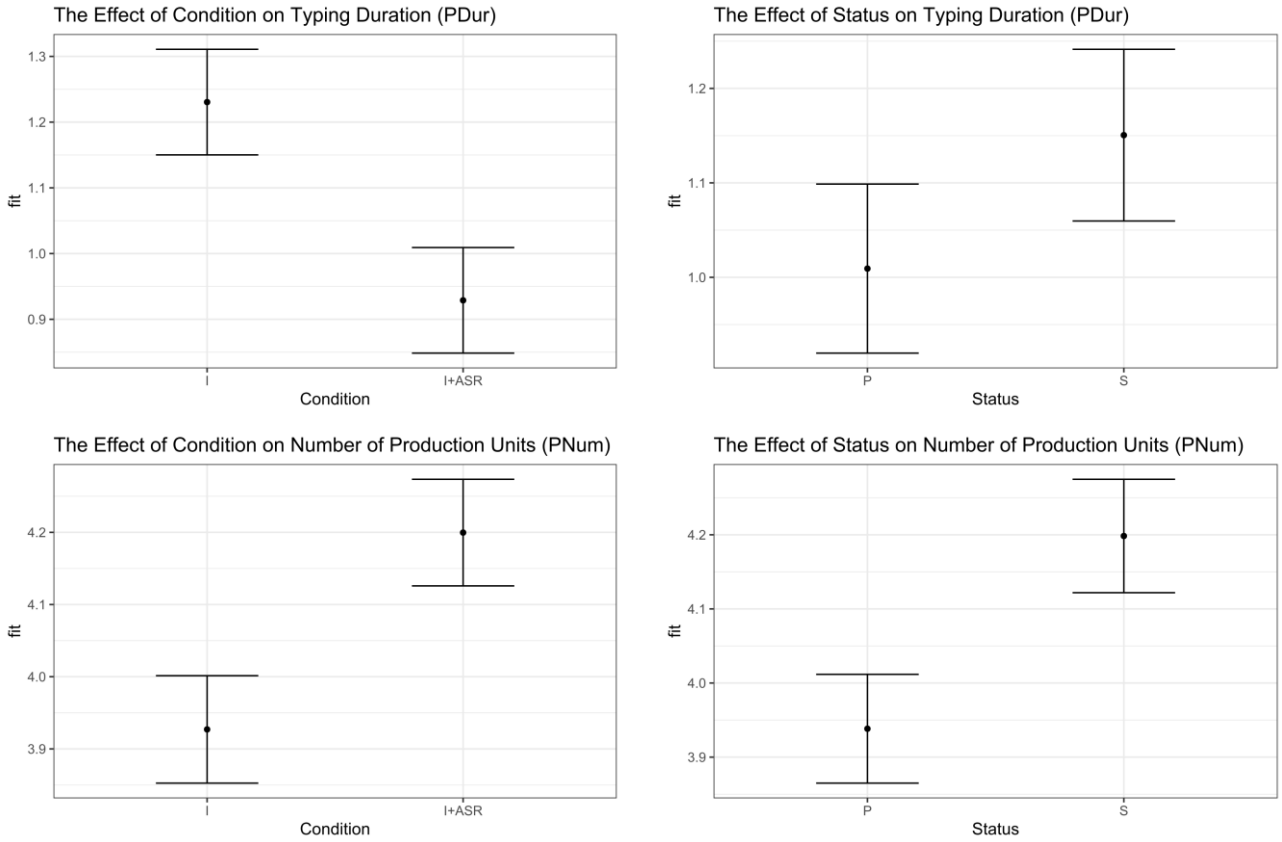
The two measures PDur and PNum were also included in models and are visualised in Figure 5. In LMM-I4, both status and condition had a significant effect on production time (PDur). For status student the effect is positive, (44 data points: β=0.1, SE=0.07, df=22, t=-2.4, p<0.05) indicating that students spent more time typing than professional subtitlers, irrespective of the kind of typing (insertion or deletion). The effect for ASR on typing time is negative (43 data points: β=-0.3, SE=0.04, df=20, t=-7, p<0.001) indicating that having an ASR script saved time spent typing, but as it did not impact the completion time, time is spent elsewhere, e.g., reading or replaying the video. Concerning production unit count (PNum) in LMM-I5, a positive effect was found for both status student (43 data

points: β=0.3, SE=0.1, df=21, t=2.7, p<0.05) and condition ASR (43 data points: β=0.3, SE=0.05, df=19, t=5.4, p<0.0001). This indicates that students interrupted their typing more often than professional subtitlers and that during the ASR condition, typing was also interrupted more often. This can be explained by the fact that before or while working with the ASR script, participants must check with the video to reject the ASR and work from scratch. Again, this will have to be looked at more closely in further analyses.

The effects plots for models LMM-I4 and I5 show that status has a positive effect for both measures; students not only take more time typing, but they also interrupt their typing more often. Regarding the effect of condition on production time, the presence of an ASR script in transcription decreases the overall typing time, but typing was interrupted more often. This suggests that working with a corrupt ASR script leads to a less efficient writing flux compared to transcribing from scratch, which professional subtitlers seem to handle better than translation students.

Figure 5

Effects for LMM-I4 and LMM-I5 in Intralingual Transcription



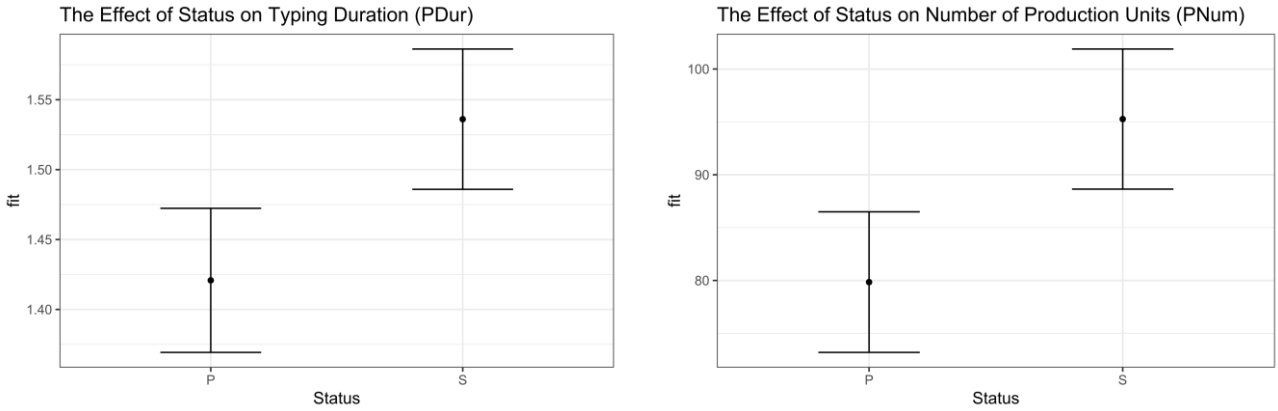
6.2.2. Translation

The effects of condition and status on technical effort in translation were tested with the same type of models as in the intralingual tasks. No significant effects were found for condition, status, or their

interactions for the direct keystroke measures, and no significant effects of condition on PDur and PNum were found, suggesting that the presence of a reference transcript does not have a significant impact on technical effort during translation. Status student, however, had a positive effect on PDur and PNum as visualized in Figure 6. In both cases, the effect is only marginally significant, and for PDur in LMM-T4 the effect was slightly smaller (66 data points: β=0.1, SE=0.06, df=23, t=1.7, p<0.1) than for PNum in LMM-T5 (64 data points: β=15, SE=7.7, df=20, t=2, p<0.06). This, again, indicates that students across all conditions spent more time typing, and typing was interrupted more often.

Figure 6

Effects for LMM-T4 and LMM-T5 in Translation



6.3. Cognitive Effort and Visual Attention

The third level, cognitive effort, is measured with gaze data, as it is linked to visual attention. This includes the mean visit duration on the ST or TT per session. Visit duration is the sum of all fixation durations during a visit from entering the AOI until leaving it. Further measures include the total visit count as well as total reading time on reference text (TrtS) and TT (TrtT). Trt is the sum of all fixation durations on the text AOI during a recording. Visual attention directed at the video was analysed with factor video replay and relative video replay duration. Due to limitations of the video player, navigation in the video was not as exact as in subtitling tools, i.e., moving per frame or second. Thus, alignment of the auditive source text was not possible. In audiovisual translation, contrary to the reading of written text, it is hard to interpret whether participants process the visual information or the synchronous audio signal at a time. This should be kept in mind when interpreting the data.

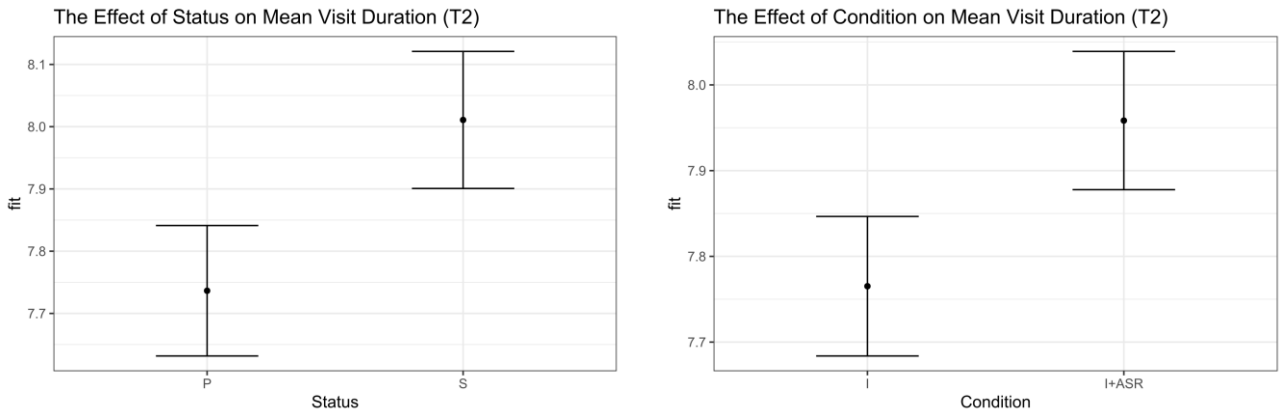
6.3.1. Intralingual Transcription

The first model LMM-I6b in Figure 7 describes the effect of status and condition on the average visit duration per session on the TT with both status student (43 data points: β=0.3, SE=0.2, df=20, t=1.8, p<0.09) and condition ASR (43 data points: β=0.2, SE=0.6, df=19, t=3.4, p<0.01) having a positive

effect. On average, students spent more time per visit on the TT, and in the ASR condition, average visit times were also longer. Given that the ASR script was in the same window as the final transcript, this effect is not surprising. An explanation for the longer visiting times of the students on the TT is that they either worked with longer passages during transcription or took more time reading the TT, be it the ASR or transcript they wrote. Professional subtitlers seem to work in shorter chunks or to be more efficient.

Figure 7

Effects for LMM-I6b in Intralingual Transcription



In LMM-I7, the effect for visit count was significant and positive for the ASR condition (44 data points: β=0.2, SE=0.04, df=17, t=3.7, p<0.01). Participants switched more often away from the TT during the ASR condition possibly to check the ASR script by replaying the video.

In LMM-I8b, the TrtT was impacted both by status and condition. Here, the DV was not log-transformed; data points were already distributed close to normal. Both effects for students (44 data points: β=1.5, SE=0.7, df=21, t=2, p<0.06) and ASR condition were positive (44 data points: β=1.7, SE=0.3, df=19, t=6, p<0.01). This makes sense as participants had to read the ASR before deciding to reject or accept it, and students, more used to written texts, spent more time reading than the professional subtitlers, who possibly also paid more attention to the video.

Participants' consulting of the video was also observed. Again, the extent to which oral and visual information was processed at a time can only be assumed. This is a problem in eyetracking and audiovisual material. The variable video replay, thus, is expressed as a factor of the video duration, i.e., the total time participants spent actually looking at the video AOI (sum of fixation durations) during video replay divided by the duration of the video scene. This factor in LMM-I9 is significantly and positively affected by status student (44 data points: β=0.4, SE=0.2, df=41, t=3, p<0.01). Students spent almost twice as much time replaying and looking at the video compared to the video length than professional subtitlers irrespective of condition. Either, professional subtitlers needed to replay the video less than students, or they were more efficient in doing so – replaying the video while typing the TT.

Further, condition ASR had a negative effect on the relative video consultation time in LMM-I10 (44 data points: β=-0.3, SE=0.2, df=23, t=-2, p<0.05). The relative time spent watching the video decreases significantly in the ASR condition, which can be explained with part of that time being spent on checking the ASR script. Participants probably played the video and listened to the audio while reading the script to check if it needed corrections. During the from scratch condition they would focus more and more often on the video since they needed to stop the video from time to time to type the transcription. Thus, inferences on the cognitive effort cannot be made directly from these measures. A closer look at the different strategies is therefore necessary but beyond the scope of this article.

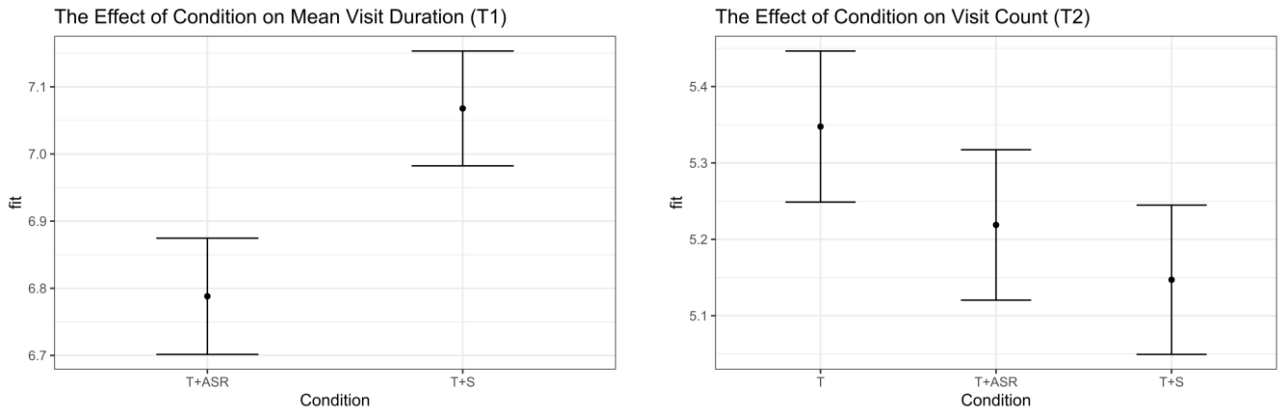
6.3.2. Translation

In LMM-T6a the DV is the average visit duration on the ST (ASR script or correct English transcript) during translation. Only the condition with a correct transcript had a significant positive effect (45 data points: β=0.3, SE=0.08, df=18, t=3.6, p<0.01). As visualised in Figure 8, on average, participants spent more time per visit in the T1 window when it contained a correct transcript suggesting that it is read more closely whereas the incomplete ASR script is, at best, consulted for individual expressions.

The effect of condition on average visit duration on TT was only marginally significant. The effect of condition on TT visit count in LMM-T7, however, is significant and negative – for ASR smaller (64 data points: β=-0.13, SE=0.05, df=34, t=-2.5, p<0.05) than for the correct English script (64 data points: β=-0.2, SE=0.05, df=34, t=-3.8, p<0.0001). Figure 8 shows a progressive negative effect of condition with increasing support on TT visit count. Thus, for the correct transcript participants switched significantly less often between windows than during translation from scratch or with ASR, which can be explained by not having to replay and stop the video as often, as the scripts serve as a form of memory aid. Here, a closer look on the different strategies on dealing with a support script is also necessary.

Figure 8

Effects for LMM-T6a and LMM-T7 in Translation

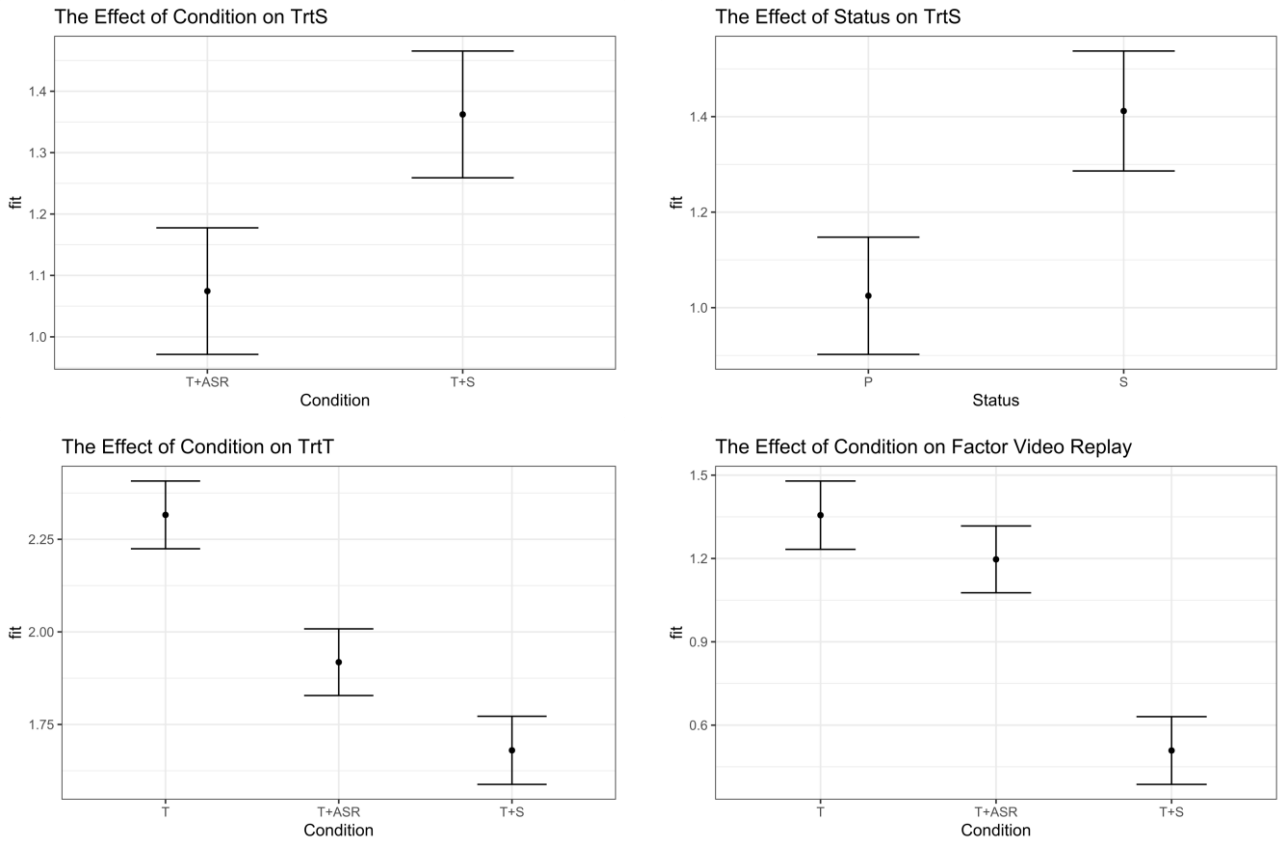


Effects on TrtS and TrtT were observed in LMM-T8a and T8b, both displayed in Figure 9. The total reading time on the T1 window (TrtS) increases significantly only for the correct transcript (44 data points: β=0.3, SE=0.1, df=21, t=3, p<0.01), and for status student (44 data points: β=0.3, SE=0.2, df=20, t=2.4, p<0.01). This means that students spent more time reading the reference script and reading times were significantly longer only for correct scripts.

Regarding the TrtT in LMM-T8b, it was affected negatively by both conditions: T+ASR (65 data points: β=-0.4, SE=0.2, df=39, t=-5.7, p<0.001) and T+S (65 data points: β=-0.6, SE=0.1, df=40, t=-9, p<0.001). Participants thus spent less time reading the TT when they had an assisting script.

Figure 9

Effects for LMM-T8a (TrtS), LMM-T8b (TrtT), and LMM-T9 (Factor Video) in Translation

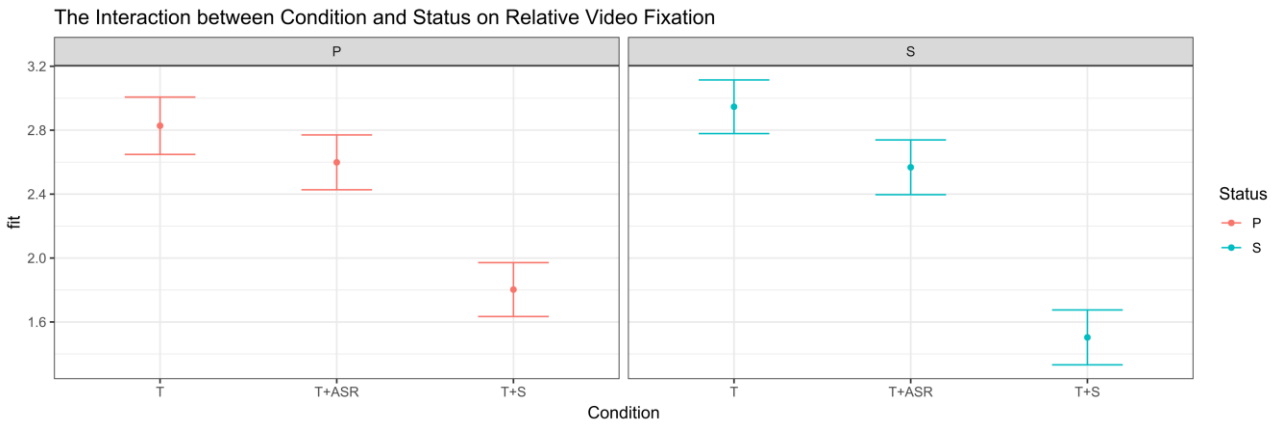


Further it was observed how much time participants spent consulting the video (see Figure 9) during translation. In LMM-T9, both conditions had a negative effect: with ASR only marginally significant (64 data points: β=-0.2, SE=0.1, df=38, t=-2, p<0.07), and for the correct script quite large and highly significant (64 data points: β=-0.8, SE=0.1, df=39, t=-10, p<0.001).

Condition also had a significant negative effect on the relative video fixation time in LMM-T10 compared to translation from scratch: with ASR (66 data points: β=-5.4, SE=1.2, df=41, t=-5, p<0.001) and with correct script (66 data points: β=-13, SE=1.1, df=40, t=-11, p<0.001). Thus, the relative video fixation time decreased with increasing support. There was also a significant interaction effect between status student and condition with correct transcript (66 data points: β=-0.4, SE=0.2, df=39, t=-2, p<0.05) as visualised in Figure 10. While all participants spent less relative time on the video with increasing support, students seem to benefit especially from the increasing quality of the reference script. This, again, can be explained with the fact that subtitlers are used to decoding audiovisual material while translation students are more used to written texts.

Figure 10

Interaction Effect for LMM-T10 between Student Status and Condition



7. Conclusion

Summarizing the explorative findings and the automatization potentials within subtitling processes, current state-of-the-art ASR seems not to come near the effects correct transcripts have in terms of temporal, technical and cognitive effort. However, since the output of these automatized processes highly depends on the quality and accessibility of training data, further tests with different quality levels of ASR are necessary. Particularly in the translation task, a correct source transcript can support the translation process tremendously on all effort levels.

In intralingual transcription, technical effort was reduced by the ASR as fewer keystrokes were performed and less time was spent typing. The relative time spent fixating the video was decreased as attention was divided between video and ASR script. These findings suggest that, while the ASR condition did not improve temporal effort, at least it impacts technical effort.

In translation, the presence of an ASR script did not decrease the temporal effort in a similar way as a correct English transcript does, which also applies to the total reading time of that transcript. Reading time was significantly longer for the correct transcript, suggesting that the ASR script was not really consulted much. Positive effects were also was found for the mean visit duration and visit count, suggesting that participants interacted much more with the correct transcript than with the ASR script. Negative effects regarding effort were found for both conditions on target text reading times and time spent watching the video. This suggests that participants tried to draw more information from the reference texts. In a next step, these measures will have to be considered regarding target text quality.

Concerning experience, differences between students and professional subtitlers were found in both tasks but not on all effort levels and not as prominent as expected. While it took students significantly longer to complete the intralingual task, there was no difference in time during translation. However,

students demonstrated higher technical effort regarding the number of production units and the overall typing time in all tasks, indicating that professional subtitlers work more efficiently. Regarding visual attention, students spent more relative time replaying the video when a source script was available and interacted more with the ASR than professional subtitlers. This finding supports the fact that especially translation students not used to work with audiovisual content benefitted from the written support. Thus, transferable subtitling skills are not necessarily (very) useful in this kind of work.

In conclusion, this study provides initial support that written assistance in AVT could be a contributor to decreased effort, if the quality of transcripts meets certain levels. The differences between professional subtitlers and translation students were not as apparent as expected. There is still much data to be analysed. In addition to investigating the different strategies applied and the TT quality, the results will have to be examined more closely with different languages and language combinations in future research. As research on AI in language technologies is constantly striving for better results, it can be expected that the proposed workflow with ASR proves to be a promising way to support subtitlers without involving expensive professional transcribers in the process.

References

Aliprandi, C., Scudellari, C., Gallucci, I., Piccinini, N., Raffaelli, M., del Pozo, A., Álvarez, A., Arzelus, H., Cassaca, R., Luis, T., & others. (2014). Automatic live subtitling: State of the art, expectations, and current trends. *Proceedings of the NAB Broadcast Engineering Conference*. Retrieved from https://doi.org/10.13140/RG.2.1.3995.3440

Álvarez, A., Arzelus, H., & Etchegoyhen, T. (2014). Towards customized automatic segmentation of subtitles. In *Advances in Speech and Language Technologies for Iberian Languages* (pp. 229-238). Springer. https://doi.org/10.1007/978-3-319-13623-3_24

Anasuya, M., & Katti, S. (2009). Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, 6, 181-205. Retrieved from https://arxiv.org/ftp/arxiv/papers/1001/1001.2267.pdf

Baños, R., & Diaz-Cintas, J. (2018). Language and translation in film: Dubbing and subtitling. In K. Malmkjær (Ed.), *The Routledge Handbook of Translation Studies and Linguistics*. London: Routledge (pp. 313-326). https://doi.org/10.4324/9781315692845

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1). https://doi.org/10.18637/jss.v067.i01

Beuchert, K. (2017). *The web of subtitling: A subtitling process model based on a mixed methods study of the Danish subtitling industry and the subtitling processes of five Danish subtitlers* (Doctoral dissertation). Department of Management, Aarhus University. Retrieved from https://www.forskningsdatabasen.dk/en/catalog/2394607615

Bywood, L., Georgakopoulou, P., & Etchegoyhen, T. (2017). Embracing the threat: Machine translation as a solution for subtitling. *Perspectives: Studies in Translation Theory and Practice*, 25(3), 492-508. https://doi.org/10.1080/0907676X.2017.1291695

Carl, M. (2012). Translog-II: A program for recording user activity data for empirical reading and writing research. *Proceedings of the 8th International Conference on Language Resources and Evaluation* (pp.153-162). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/index.html

Carl, M., Schaeffer, M., & Bangalore, S. (2016). The CRITT translation process research database. In M. Carl, S. Bangalore, & M. Schaeffer (Eds.), *New directions in empirical translation process research—Exploring the CRITT TPR-DB* (pp. 13-56). Springer. https://doi.org/10.1007/978-3-319-20358-4

Del Pozo, A., Aliprandi, C., Álvarez, A., Mendes, C., Neto, J. P., Paulo, S., Piccinini, N., & Raffaelli, M. (2014). SAVAS: Collecting, annotating, and sharing audiovisual language resources for automatic subtitling. *Proceedings of the 9th International Conference on Language Resources and Evaluation* (pp. 432-436). Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/index.html

Díaz-Cintas, J. (2013). The technology turn in subtitling. *Proceedings from the 5th International Maastricht - Łódź Duo Colloquium on Translation and Meaning, Part 9* (pp. 119-132). Retrieved from https://www.researchgate.net/profile/Jorge_Diaz-Cintas/publication/314262159_The_technology_turn_in_subtitling/links/58bea66792851c971449f6bc/The-technology-turn-in-subtitling.pdf

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

Díaz-Cintas, J., & Massidda, S. (2019). Technological advances in audiovisual translation. In Minako O'Hagan (Ed.), *The Routledge Handbook of Translation and Technology*. (pp. 255-270) London: Routledge. https://doi.org/10.4324/9781315311258

Georgakopoulou, P. (2019). Template files: The holy grail of subtitling. *Journal of Audiovisual Translation*, 2(2), 137-60. Retrieved from http://www.jatjournal.org/index.php/jat/article/view/84

Hvelplund, K. T. (2017). Eye tracking and the process of dubbing translation. In J. Díaz Cintas & K. Nikoli (Eds.), *Fast-forwarding with audiovisual translation* (pp. 110-124). Multilingual Matters. https://doi.org/10.21832/9781783099375-010

Just, Marcel Adam, and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329-54.

Kordoni, V., Birch, L., Buliga, I., Cholakov, K., Egg, M., Gaspari, F., Georgakopoulou, Y., Gialama, M., Hendrickx, I., Jermol, M., & others. (2016). TraMOOC (Translation for Massive Open Online Courses): Providing reliable MT for MOOCs. *Proceedings of the 19th annual conference of the European Association for Machine Translation*. Retrieved from https://pure.uvt.nl/ws/portalfiles/portal/17800956/tramooc.pdf

Krings, H. P. (2001). *Repairing texts: Empirical investigations of machine translation post-editing processes* (G. S. Koby & G. M. Shreve Trans.). Kent, OH: Kent State University Press. (Original work published 1997).

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2019*). lmerTest: Tests in linear mixed effects models*. https://CRAN.R-project.org/package=lmerTest

Matamala, A., Romero-Fresco, P., & Daniluk, L. (2017). The use of respeaking for the transcription of non-fictional genres: An exploratory study. *InTRAlinea: Online Translation Journal*, 19. http://www.intralinea.org/archive/article/2262

Media Consulting Group & Peacefulfish. (2007). *Study on dubbing and subtitling needs and practices in the European audiovisual industry* [PDF document]. Retrieved from http://www.lt-innovate.org/sites/default/files/documents/1342-Study%20on%20dubbing%20and%20subtitling%20needs%20and%20practices%20in%20the%20European%20audiovisual%20industry%20(2008).pdf

Melero, M., Oliver, A., & Badia, T. (2006). Automatic multilingual subtitling in the eTITLE project. *Proceedings from Translating and the Computer 28* (pp. 1-18). Retrieved from http://www.mt-archive.info/Aslib-2006-Melero.pdf

Orrego-Carmona, D., Dutka, Lukasz, & 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. Retrieved from https://jostrans.org/issue30/art_orrego-carmona_et_al.php

Quintas, L. C. (2017). Towards a hybrid Intralinguistic subtitling tool: Miro translate. *Proceedings from Translating and the Computer 39* (pp. 1-18). Retrieved from https://biblio.ugent.be/download/8549926/8549927.pdf#page=10

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raats, T., Evens, T., & Ruelens, S. (2016). Challenges for sustaining local audio-visual ecosystems: Analysis of financing and production of domestic TV fiction in small media markets. *The Journal of Popular Television*, 4(1), 129-147. https://doi.org/10.1386/jptv.4.1.129_1

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62(8), 1457-1506. https://doi.org/10.1080/17470210902816461

Schneeberger, A. (2019). *Audiovisual media services in Europe. Market insights*. European Audiovisual Observatory. Retrieved from https://rm.coe.int/audiovisual-media-services-in-europe-market-insights/16809816d1

Silvestre Cerdà, J. A., Del Agua Teba, M. A., Garcés Díaz-Munío, G. V., Gascó Mora, G., Giménez Pastor, A., Martínez-Villaronga, A. A., Pérez González de Martos, A. M., Sánchez-Cortina, I., Serrano Martínez-Santos, N., Spencer, R. N., & others. (2012). Translectures. *Proceedings from IberSPEECH '12* (pp. 345-351). Retrieved from http://hdl.handle.net/10251/37290

Tardel, A. (Forthcoming). Measuring effort in subprocesses of subtitling. The case of post-editing via pivot language. In M. Carl (Ed.), *Recent advances in empirical translation process research.* https://sites.google.com/site/centretranslationinnovation/conferences-workshops/memento-ws-2019

Valor Miró, J. D., Spencer, R. N., Pérez González de Martos, A., Garcés Díaz-Munío, G., Turró, C., Civera, J., & Juan, A. (2014). Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. *Open Learning: The Journal of Open, Distance and e-Learning*, 29(1), 72-85. https://doi.org/10.1080/02680513.2014.909722

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer. Retrieved from https://ggplot2.tidyverse.org