Comprehension of Subtitles from Re-Translating Simultaneous Speech Translation

Anonymous ACL-IJCNLP submission

Abstract

In simultaneous speech translation, one can vary the size of the output window, system latency and sometimes the allowed level of rewriting. The effect of these properties on readability and comprehensibility has not been tested with modern neural translation systems. In this work, we propose an evaluation method and investigate the effects on comprehension and user preferences. It is a pilot study with 14 users on 2 hours of German documentaries or speeches with online translations into Czech. We collect continuous feedback and answers on factual questions. Our results show that the subtitling layout or flicker have a little effect on comprehension, in contrast to machine translation itself and individual competence. Other results show that users with a limited knowledge of the source language have different preferences to stability and latency than the users with zero knowledge. The results are statistically insignificant, however, we show that our method works and can be reproduced in larger volume.

1 Introduction

Simultaneous speech translation is a technology that assists users to understand and follow a speech in a foreign language in real-time. The users may need such an assistance because of limited knowledge of the source language, the speaker's non-native accent, or the topic and vocabulary. The technology can be used for the target languages, for which human interpretation is unavailable, e.g. due to capacity reasons.

The candidate systems for simultaneous speech translation differ in quality of translation, latency and the approach to stability. Some are streaming, only adding more words (Grissom II et al., 2014; Gu et al., 2017; Arivazhagan et al., 2019; Press and Smith, 2018; Xiong et al., 2019; Ma et al., 2019; Zheng et al., 2019), some allow re-translation as more input arrives (Müller et al., 2016b; Niehues et al., 2016; Dessloch et al., 2018; Niehues et al., 2018; Arivazhagan et al., 2020). Finally, subtitle presentation options (size of subtitling window, layout, allowed reading time, font size, etc.) also affect users' impression. The re-translating speech-to-text translation systems can offer lower latency by producing partial text hypotheses, which are however often withdrawn and replaced by new, more accurate versions. The combination of the re-translating approach and limited space for subtitles is challenging because of "flicker" by which we mean all the re-translations of the text that a user is reading at the moment, has already read, or that has been scrolled away. In this case, the subtitling options impact the reading comfort and delay and may affect the general usability.

The evaluation of the traditional, text-to-text machine translation (MT) has been researched for many years (see e.g. Han, 2018 or developments and discussion within the series of WMT, Barrault et al., 2020). It targets only the translation quality.

Simultaneous speech translation evaluation faces new challenges: simultaneity, latency, and readability to humans. Evaluating only selected aspects in isolation is reasonable (as quality in Elbayad et al., 2020), however, a complete evaluation must be end-to-end, from sound acquisition to subtitling and testing whether the users received the information.

We propose a method for human evaluation of simultaneous translation on simulated live events. We focus on the evaluation of subtitling layouts and measuring comprehension effectively. We demonstrate our method on 14 users and 15 video or audio documents (115 minutes in total) in German with one online translation system into Czech. We collect the users' feedback on the quality of subtitles during watching, and ask them to answer questions on information from the video to measure their

comprehension.

We have no prior estimate on the statistical significance of results with the limited number of participants and documents. In this pilot study, we test the significance and give the estimate for further, more extensive studies.

Our results showed that our speech translation system preserves on average 80% of information from the source, when used in offline mode, i.e. when the user has unlimited time to browse the translation. An average single person is able to find around 33% of information in online mode. Next, we found an optimal subtitling layout, and found that its difference from a suboptimal, but reasonable layout is small and insignificant. Finally, we tested if the evaluation can be simplified by using judges with a knowledge of the source language without comprehension questionnaires.

2 Related Work

Hamon et al. (2009) propose user evaluation of speech-to-speech simultaneous translation. To test the adequacy and intelligibility, they prepared questionnaires with factual questions from the source speech. The judges listened either to the interpreter, or the machine, and answered the questions. They evaluated the offline mode, the judges were allowed to stop and replay the audio while answering. This way the authors measured the comprehension loss caused by the automatic translation or interpretation. Each sample was processed by multiple judges, to eliminate human errors. Fluency was assessed by the judges on a scale.

Macháček and Bojar (2020) propose a technique for collecting continuous user rating while the user watches video and simultaneous subtitles. The user is asked to express the satisfaction with the subtitles at any moment by pressing one of four buttons as the rating changes.

Müller et al. (2016a) analyzed the feedback from foreign students using KIT Lecture Translator within two semesters. Such a long-term and informal evaluation differs considerably from judging in controlled conditions. On one hand, it summarizes the real-life situation with all the variables and corner cases that a lab test could only approximate or omit. On the other hand, the users may not be motivated to give the feedback, and can give only personal opinions that may be biased. This way it is also difficult to compare multiple system candidates.

3 Evaluation Campaign

In our evaluation, we simulate live events at which participants need assistance with understanding the spoken language. We prepared a web application presenting video or audio documents equipped with live subtitles. The judges see each document for their first time, only once, with source sound and without interruptions, to simulate the live setting. While watching, they press buttons to indicate their current satisfaction with the subtitles. Afterwards, they fill a questionnaire with comprehension and summary questions. We distribute different versions of subtitling setups among the judges for contrastive analysis.

The source and target languages in our study are German and Czech, respectively. This is an interesting example of two neighbouring countries, distinct language families and yet a relatively well studied pair with sufficient direct training data.

3.1 Translation System

We use the ASR system originally prepared for German lectures (Cho et al., 2013). It is a hybrid HMM-DNN model emitting partial hypotheses in real time, and correcting them as more context is available. The same system was used also by KIT Lecture Translator (Müller et al., 2016b).

The system is connected in a cascade with a tool for removing disfluencies and inserting punctuations (Cho et al., 2012), and with a German–Czech NMT system.

The machine translation is trained on 8M sentence pairs from Europarl and Open Subtitles (Koehn, 2005; Lison and Tiedemann, 2016), the only public parallel corpora of German and Czech, and validated on newstest. The Transformer-based (Vaswani et al., 2017) system runs in Marian (Junczys-Dowmunt et al., 2018) and reaches 18.8 cased BLEU on WMT newstest-2019.

Despite the translations are pre-recorded and only played back in our simulated setup, we ensured we keep the original timing as emitted by the online speech translation system.

3.2 Selection of Documents

We selected German videos or audio resources that fulfilled following conditions: 1) Length 5 to 10 minutes (with few exceptions). 2) The translations had to be of a sufficient quality. Based on a manual check, we discarded several candidate documents: a math lecture and broadcast news due

domain	type	docs.	duration	description
EP	TP	3	18:08	From European Parliament
DG	TP	3	17:34	From DG SCIC repository for interpretation training
Mock Int	A	3	27:52	From a mock interpreted conference at interpretation school
Maus	V	2	14:43	Educative videos for children
DW	A	2	18:48	Audio for intermediate learners of German
Dinge	V	2	16:09	Educative video for teenagers and grown-ups
All		15	114:52	

Table 1: Summary of domains of selected documents. "Type" distinguishes audio only (A), talking person only (TP) and video (V) with illustrative or informative content. Duration is reported in minutes and seconds.

to many mistranslated technical terms and named entities. Another group of documents was mistranslated and discarded because they were not long-form speeches, but isolated utterances with long pauses. 3) Informative content. We intend to measure adequacy and comprehension by asking the judges complementary questions. We thus excluded the documents where the speaker is not giving information by speech, but uses mostly paralinguistic means, e.g. singing, poetry, or non-verbal communication. 4) Non-technicality. We expect the judges answer in several plain words in their mother tongue. They may lack knowledge of any specialized vocabulary.

We selected audios, videos with informative or illustrative content, and videos of talking persons, to compare user feedback for these types of documents.

Table 1 summarizes the selected documents.

3.3 Questionnaires

We decided to use direct factual questions in our study, instead of yes/no questions to exclude guessing. We asked a Czech teacher of German to prepare the questions and an answer key from the original German documents, regardless of the machine translation. The teacher wrote the questions in Czech, and was instructed to prepare one question from every 30 seconds of the stream and distribute them evenly, if possible. The questions had to be answerable only after listening to the document, and not from the general knowledge. The complexity of the questions was targeted on the level that an ordinary high-school student could answer after listening to the source document once, if the student would not have any obstacles in understanding German. To reduce the effect of limited memory, the judges had an option in the questionnaire to indicate they knew the answer but forgot it. Furthermore, they had to fill, from which source they knew the answer: from the subtitles, from the speech, from an image on the video, or from their previous knowledge.

level	count	group	total
0	5	non-German speaking	10
A1	5		
A2	1	German speaking	4
B1	2		
B2	1		
All			14

Table 2: The judges by their German proficiency levels on CEFR scale and their distribution to groups.

After the factual questions, all the questionnaires had a common part where we asked the judges on their general impression of translation fluency, adequacy, stability and latency, overall quality, video watching comfort, and a summary comment. Each judge spent in total 2 hours on watching and 3 hours on the questionnaires.

Finally, we evaluated the factual questions manually against the key, rating them at three levels: correct, incorrect, and partially correct.

3.4 Judges

We selected 14 native Czech judges. Their self-reported knowledge of German had to be between zero and B2 on the CEFR[1] scale, to ensure they need some level of assistance with understanding German. We also ensured they do not have knowledge of any other language which could help them understanding German. The summary of their proficiency in German is in Table 2. For further analyses in our study, we divided them into two groups. For brevity further in the paper, we denote the 10 judges

[1]Common European Framework of Reference for Languages

Figure 1: A detail of the default layout with the video document "Dinge Erklärt: Impfen..." (https://youtu.be/4E0dwFS72gk). The video is at the top, below are two lines of subtitles in Czech, followed by buttons for the continuous quality rating. The button labels are "1 = worse", "2 = average", "3 = OK", "0 = I do not understand at all". The order 1, 2, 3, 0 matches the keyboard layout; users were encouraged to use keyboard shortcuts.

with zero or A1 level (beginners) as "non-German speaking", and the others as "German speaking". Because we have a small amount of German speaking ones, we do not classify them in more detail.

The judges were paid for participation in the study. They watched the videos at their homes on their own devices. They were asked to customize their screen resolution and eye-screen distance to suit their comfort.

3.5 Subtitler: Subtitle Presentation

The Subtitler is our implementation of the algorithm by Macháček and Bojar (2020) extended with automatic adaptive reading speed in addition to the "flicker" parameter as defined in the paper. The speed varies between 10 and 25 characters per second depending on the current size of the incoming buffer. The default font size is 4.8 mm. The default subtitling window is 2 lines high and 163 mm wide. By default, we use the maximum flicker and the lowest delay (presenting all translation hypotheses, not filtering out the partial and possibly unstable ones), no colour highlighting, and smooth slide-up animation while scrolling. The example of the setup can be seen in Figure 1.

With the default subtitling window, 90% of the words in the test documents are finalized in subtitles at most 3 seconds after translation. In 99%, it is at most 7 seconds.

Type	w. avg±std	t-test
Offline+voting	0.81±0.11	
Offline	0.59±0.16	***
Online, without flicker	0.36±0.16	***
Online, flicker, top layout	0.33±0.13	
Online, flicker, least preferred	0.31±0.16	

Table 3: Comprehension scores on all documents and judges. The average weighted by number of questions in document. *** denote the statistically significant difference (p-value<0.01) between the current and previous line.

4 Results

4.1 Comprehension

In our study, we assume that comprehension can be assessed as a proportion of correctly answered questions. We assume the following model: A person without any language barrier and with non-restricted access to the document during answering the questionnaire can answer all questions correctly. With a language barrier and offline machine translation (unlimited perusal of the document while answering), some information may be lost in machine translation. More information is lost with one-shot access to online machine translation because of forgetting and temporal inattention. Some more information may be lost because of flicker, and some more because of suboptimal subtitling layout.

Our results confirmed the assumed hierarchy of comprehension levels. Moreover, we noticed that even the judges with offline MT gave inconsistent answers. When we combined them and counted as correct if at least one was correct, they achieved higher scores. We explain it by insufficient attention.

Table 3 summarizes the results on all documents. We measured that on average, 81% of information was preserved by machine translation (Offline+voting, i.e. one of two judges answered correctly). A single judge could find 59% of information (Offline). In an oracle experiment without flicker, when the machine translation gives the final hypotheses with the timing of the partial ones (i.e. as if it knew the best translation of the upcoming sentence), a single judge could answer 36%. In real setup with flicker and the most preferred subtitling layout (Online, flicker, top layout), 33% information was found, and 31% with less preferred. The standard deviation is between 11 and 16%.

We found statistically significant difference (two-sided t-test) between offline MT with voting and

	German≥A2 # avg±std	German<A2 # avg±std	t-test
flicker	3 **0.59±0.15**	10 0.30±0.15	$p < 0.05$
no flicker	4 0.40±0.06	10 **0.34±0.07**	insig.
t-**test**	$p < 0.10522$	insignificant	

Table 4: Comprehension scores on two documents on a setup with and without flicker, as rated by judges whose German competence is between A2 and B2 on CEFR scale (elementary to upper intermediate), or below A2 (zero or beginner). Number of samples is denoted as "#", higher scores bolded.

without it, and between offline MT and online. The difference caused by flicker or layout was insignificant.

4.2 Preferences by Language Skills

We assume that the user behaviour differs by knowledge of the source language. The users with zero knowledge read all subtitles all the time and do not pay attention to the speech. They do not mind large latency, but demand high quality translation, and comfortable reading without flicker. On the other hand, the users with a limited, but nonzero knowledge of the source language listen to the speech, try to understand on their own, and look at the subtitles only occasionally, when they are temporarily uncertain or need assistance with an unfamiliar word. They need low latency, and do not mind slightly lower quality.

To empirically test our hypothesis, we prepared two setups: With flicker, the subtitles are presented immediately as available, but with frequent rewriting, which discomforts the reader. For comparison without flicker, we present only the final translations without rewriting, but with a large latency. We selected two videos and distributed these setups uniformly between German speaking and non-German speaking judges.

The results of comprehension are in Table 4. It shows that German-speaking users achieve higher comprehension with flicker than without. We consider the difference as close to statistically significant (p-value < 0.10522), although we had only 4 and 10 German and non-German speaking judges, respectively. The non-German speakers understood better without flicker (34% vs 30%), but this difference is statistically insignificant. The other types of feedback (weighted average of continuous rating and the overall rating at the end of questionnaire) confirm the trend of comprehension, but have larger variance and the differences are insignificant.

		Side	Below
Final rating	audio	3 2.00 ±0.82	6 2.00 ±0.82
	talking	4 2.25 ±0.83	3 **2.67 ±0.94**
	video	1 1.00 ±0.00	1 1.00 ±0.00
	sum, avg	8 2.00 ±0.87	10 **2.10 ±0.94**
Comprehension	audio	3 **0.27 ±0.13**	6 0.21 ±0.13
	talking	4 0.22 ±0.12	3 **0.28 ±0.26**
	video	1 0.18 ±0.00	1 **0.33 ±0.00**
	sum, avg	8 0.23 ±0.12	10 **0.24 ±0.18**
Avg. cont. rating	audio	3 **1.18 ±0.76**	6 0.76 ±0.54
	talking	4 1.20 ±0.79	3 **1.76 ±0.47**
	video	1 0.23 ±0.00	1 **0.77 ±0.00**
	sum, avg	8 **1.07 ±0.79**	10 1.06 ±0.67
Watching comfort	talking	4 2.75 ±0.83	3 **3.00 ±0.82**
	video	1 2.00 ±0.00	1 **3.00 ±0.00**
	sum, avg	5 2.60 ±0.80	4 **3.00 ±0.71**

Table 5: Results of the contrastive experiments of the non-German speaking judges for side vs below layout. The three numbers in each row and cell are the number of experiments, average and standard deviation. The higher score, the better. Comprehension rate is between 0 and 1, average continuous rating is between 0 and 3, the others on a discrete scale 1 to 5. Higher score in each row bolded.

4.3 Subtitling Layout

We analyzed effects of distinct subtitling features by contrastive experiments differing only at one feature. We distributed them randomly among the judges, regardless of their German skills. We can draw conclusions only on non-German speaking judges due to insufficient number of observations for the German-speaking group.

In all cases, the results show a slight insignificant preference towards one variant of the feature in all three types of feedback (comprehension, weighted average of continuous rating, and overall rating at the end of video).

4.3.1 Side vs Below

For videos and videos with a talking person, we consider two locations for the subtitle window: on the left side of the video, or below. The side window can be high but narrow (17 lines of 60 mm width, to match the height of the video), while the window underneath is short and wide (2 lines of 163 mm width). The first is more comfortable for reading, the latter for watching video.

The results are in Table 5. "Final rating" and "Watching comfort" summarize the responses in the final section of the questionnaire, where judges answered on a discrete scale 1 (worst) to 5 (best). "Comprehension" and "Average continuous rating" are, as above, results from correctness of answers and from the feedback button clicks, resp. The

		Below	Overlay
Final rating	talking	9 2.33 ±1.05	9 **2.78 ±1.13**
	video	5 1.40 ±0.80	8 **2.38 ±0.86**
	sum, avg	14 2.00 ±1.07	17 **2.59 ±1.03**
Compre-hension	talking	9 0.29 ±0.25	9 **0.39 ±0.20**
	video	5 0.26 ±0.14	8 **0.37 ±0.11**
	sum, avg	14 0.28 ±0.21	17 **0.38 ±0.17**
Avg. cont. rating	talking	9 1.65 ±0.52	9 1.65 ±0.99
	video	5 1.11 ±0.50	8 **1.15 ±0.77**
	sum, avg	14 **1.47 ±0.57**	17 1.42 ±0.93
Watching comfort	talking	9 3.43 ±0.73	9 **4.11 ±0.74**
	video	5 2.20 ±1.60	8 **3.00 ±1.00**
	sum, avg	14 2.92 ±1.32	17 **3.59 ±1.03**

Table 6: Results of the experiments on "overlay" vs "below" layout, for non-German speaking judges. Description of numbers and ratings as in Table 5.

Size [lines,mm width]	18×250 ("Large")	
Highlighting	No	Yes
Final rating	14 2.93 ±0.80	13 **3.31 ±1.14**
Comprehension	14 0.25 ±0.15	13 **0.30 ±0.12**
Avg. cont. rating	14 1.32 ±0.82	13 **1.42 ±0.74**

Size [lines,mm width]	5×200 ("Medium")	
Highlighting	No	Yes
Final rating	2 2.50 ±0.50	1 **4.00 ±0.00**
Comprehension	2 **0.44 ±0.18**	1 0.39 ±0.00
Avg. cont. rating	2 **2.19 ±0.50**	1 2.12 ±0.00

Table 7: Results of highlighting experiments on audio documents. Description of numbers as in Table 5.

results show statistically insignificant difference in all measures. There is a slight overall preference for the layout "below", except audio-only documents.

4.3.2 Overlay vs Below

The subtitling window can be placed over the video, as in films, or below. In the first case, the subtitles possibly hide an informative image content, in the latter case, there is a larger distance between the image and the subtitles. The results on non-German speaking judges are insignificantly in favor of overlay, see Table 6.

4.3.3 Highlighting Flicker Status

The underlying rewriting speech translation system distinguishes three levels of status for segments (automatically identified sentences): "Finalized" segments means no further changes are possible. "Completed" segments are sentences which received a punctuation mark. They can be changed by a new update and the prediction of the punctuation may also change or disappear. They usually flicker once in several seconds. "Expected" segments are incomplete sentences, to which new translated words are still appended. They flicker several times per second.

Size [lines,mm width]		2×163	5×200
Final rating	audio	10 1.80 ±0.87	8 **2.75 ±0.97**
	talking	9 2.33 ±1.05	5 **2.80 ±1.60**
	video	5 1.40 ±0.80	3 **2.33 ±0.47**
	sum, avg	24 1.92 ±1.00	16 **2.69 ±1.16**
Compre-hension	audio	10 0.25 ±0.15	8 **0.31 ±0.15**
	talking	9 0.29 ±0.25	5 **0.40 ±0.21**
	video	5 0.26 ±0.14	3 **0.28 ±0.05**
	sum, avg	24 0.26 ±0.19	16 **0.33 ±0.16**
Avg. cont. rating	audio	10 0.90 ±0.71	8 **1.66 ±0.95**
	talking	9 **1.65 ±0.52**	5 1.09 ±0.78
	video	5 1.11 ±0.50	3 **1.35 ±0.31**
	sum, avg	22 1.21 ±0.70	16 **1.42 ±0.85**
Watching comfort	talking	7 **3.43 ±0.73**	5 2.80 ±0.98
	video	5 2.20 ±1.60	3 **2.33 ±1.25**
	sum, avg	12 **2.92 ±1.32**	8 2.62 ±1.11

Size [lines,mm width]		18×250	5×200
Final rating	audio	11 **2.91 ±0.79**	8 2.75 ±0.97
Comprehension	audio	11 0.23 ±0.14	8 **0.31 ±0.15**
Avg. cont. rat.	audio	11 1.50 ±0.79	8 **1.66 ±0.95**

Table 8: Results of the experiments with subtitling window. Descriptions as in Table 5.

It is a user interface question if the status of the segments should be indicated by highlighting, or if this piece of information would be rather disturbing. We experimented only with colouring text background in large and medium subtitling window for audio-only documents.

Our experiments show that the judges prefer highlighting flicker status in the large window. For the medium window, this inclination is less clear, see Table 7.

4.3.4 Size of Subtitling Window

The subtitling window can be of any size. If the window is short and narrow, there is a short gap between an image and subtitles, which simplifies focus switching. On the other hand, a small window contains short history, so the user can miss translation content if it disappears while paying attention to the video. A small window may also cause a long subtitling delay if the translation was updated in scrolled away part of text, so that Subtitler has to return and repeat it (a very disturbing "reset"). With a large window, there is a larger distance between the end of subtitles and the image. The content stays longer, but it is more complicated to find a place where the user stopped reading before the last focus switch.

Depending on spatial constraints, it is always recommended to use as large window as possible, especially for documents without visual information, where focus switching between an image and

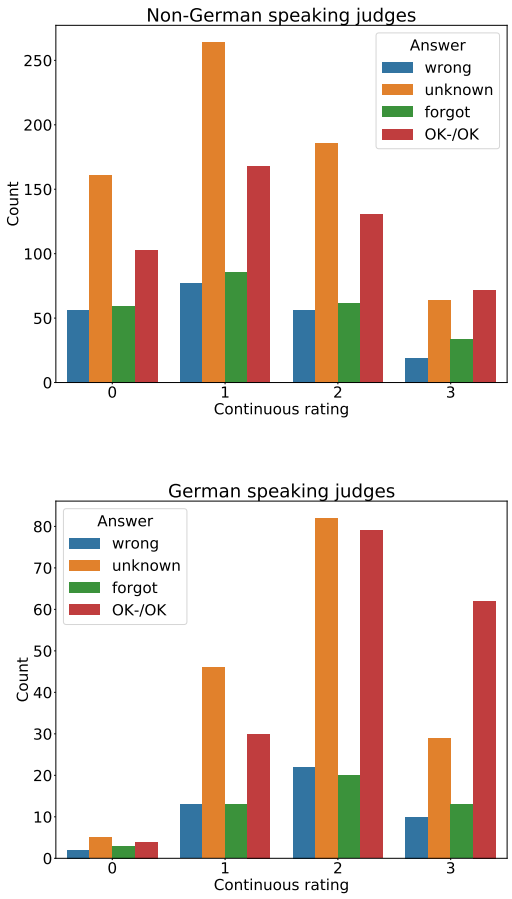

Figure 2: The distribution of the continuous rating and results of answers for non-German (upper) and German speaking (lower) judges.

subtitles is not expected. We tested two pairs of sizes on the same documents. The results are in Table 8. As we expected, the window with 5 lines was rated insignificantly better than with 2 lines, but the 2-line was more comfortable for watching. The judges rated it with average 2.92 in final section of the questionnaire, while the 5-line average was 2.62.

For an audio-only document, we also tested the large (18 lines) vs. medium (5 lines) window, observing users' reported preference for the large one but slightly higher comprehension and continuous feedback for the medium one, see the lower part of Table 8.

4.4 Relating Comprehension and Continuous Rating

We collected continuous rating of the overall quality of subtitles at given times, with four levels, where 0 means the worst and 3 the best. For every

	χ^2-test p-values			
answers	Non-G. sp. j.		Germ. sp. judges	
wrong	0.53	insig.	0.81	insig.
unknown	0.28	insig.	**0.09**	sign. $p < 0.1$
forgot	0.69	insig.	0.61	insig.
OK/OK-	0.12	insig.	**0.03**	sign. $p < 0.05$

Table 9: The results of χ^2-test for statistical significance of the independence of the distribution of continuous ratings and answer correctness.

comprehension question, we know the time when the necessary piece of information is uttered in the source speech document. Based on this timing information, we can relate comprehension and the reported continuous feedback. In Figure 2, we plot the number of Continuous rating button clicks divided according to whether the information at that time was understood acceptably ("OK/OK-"), spotted but forgotten ("forgot"), missed by the user ("unknown"), or misunderstood ("wrong"). This data aggregates observations for all documents and all setups excluding the offline MT and the oracle online MT without flicker.

We use the χ^2-test to measure whether the distribution of answer results and continuous rating are independent or not. The results are in Table 9. For the non-German speaking judges, the distributions are independent, while for the German speaking there is a statistically significant dependence between unknown answers and ratings, and correct answers and ratings. It means that if we know the ratings of the German speaking judges, we can predict their comprehension with a higher precision than without it. This observation could be used as the basis for a less time-consuming evaluation, e.g. when several translation systems need to be compared. Judges with elementary to upper intermediate knowledge of the source language could only watch the subtitles and provide continuous feedback, instead of the comprehension questions. The questions are laborious to both prepare and answer.

Forgetting and wrong answers are found to be independent on the continuous feedback. It is possible that the wrong answers are caused by inadequacies in the machine translation that non-German speakers can not observe, which are distributed uniformly regardless the flicker, latency or fluency.

From the χ^2 test results, we conclude that for the non-German speaking judges, their comprehen-

sion is probably independent of their continuous rating, because they have no competence for rating the adequacy. Their ratings are based only on fluency, readability and flicker. The German-speaking judges probably included the adequacy factor into the rating, which the non-German speakers could not do. This fact could be used in the future works. The judges could be used for comparison of multiple translation candidates. The judges who speak the source language could assess the adequacy only by the continuous rating without the need for questionnaires, which are laborious to prepare, answer and evaluate. The non-German speaking judges could skip the continuous rating and only fill out the questionnaire for adequacy.

5 Scalability

The evaluation method described in this paper requires manual work to select the documents, prepare, fill and evaluate the questionnaires. The amount of work is feasible in small number of documents and judges, but the results are insignificant. Re-scaling to large volumes may be costly. Therefore, in this section we propose ways to reduce the manual work in future evaluations.

It is advisable to target only on the documents, on which the speech translation achieves sufficient quality, because the users' impression will be equally bad with low-quality translations. The quality can be estimated by automatic MT metrics (e.g. BLEU, METEOR, etc.), if the reference translations are available.

We hypothesize that the questionnaires can be avoided, if future works confirm correlation of continuous rating of bilingual judges with adequacy. To measure the correlation and limits of significancy, experiments with large amounts of manual work are necessary, similarly as when finding the evidence for correlation of BLEU to human judgements (Reiter, 2018).

6 Conclusion

We proposed a method for end-to-end user evaluation of simultaneous speech translation, relying on users' continuous feedback and a follow-up questionnaire. The method can be used for measuring comprehension and evaluating subtitling parameters. We test the method in an evaluation campaign using 14 judges and 115 minutes of video and audio documents. Each of the judges spent 2 hours watching the documents and 3 hours answering the questionnaires. We observed that with the judges knowing the source language, it could be possible to omit the questionnaires because they seem to be able to assess adequacy in continuous rating.

The most preferred subtitling parameters are two lines of subtitles placed over the video, if the video has informative content. In case of video with a talking person or audio document, the most preferable is a large subtitling window with colour indication of whether the segment is final or still can change.

The users with a knowledge of the source language prefer low latency for sake of stability, while the users without language knowledge have no preference.

We did not find a statistically significant evidence on the impact of the differences in subtitling parameters to comprehension. We hypothesize that if the parameters are reasonable and do not cause a large delay, then the effect is close to zero. The largest effect on comprehension can be attributed to the individual competence and machine translation.

We successfully tested the method on limited number of participants and documents, and got statistically insignificant results. We conclude that our work may be used for an estimate of significance for further, more extensive studies.

References

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages

1–55, Online. Association for Computational Linguistics.

Eunah Cho, C. Fügen, T. Hermann, K. Kilgour, Mohammed Mediani, C. Mohr, J. Niehues, Kay Rottmann, C. Saam, Sebastian Stüker, and A. Waibel. 2013. A real-world system for simultaneous translation of german lectures. pages 3473–3477.

Eunah Cho, J. Niehues, and Alexander H. Waibel. 2012. Segmentation and punctuation prediction in speech language translation using a monolingual translation system. In *IWSLT*.

Florian Dessloch, Thanh-Le Ha, Markus Müller, Jan Niehues, Thai-Son Nguyen, Ngoc-Quan Pham, Elizabeth Salesky, Matthias Sperber, Sebastian Stüker, Thomas Zenkel, and Alexander Waibel. 2018. KIT lecture translator: Multilingual speech translation with one-shot learning. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 89–93, Santa Fe, New Mexico. Association for Computational Linguistics.

Maha Elbayad, Michael Ustaszewski, Emmanuelle Esperança-Rodier, Francis Brunet-Manquat, Jakob Verbeek, and Laurent Besacier. 2020. Online versus offline NMT quality: An in-depth analysis on English-German and German-English. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5047–5058, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1342–1352, Doha, Qatar. Association for Computational Linguistics.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

Olivier Hamon, Christian Fügen, Djamel Mostefa, Victoria Arranz, Muntsin Kolss, Alex Waibel, and Khalid Choukri. 2009. End-to-end evaluation in simultaneous translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 345–353, Athens, Greece. Association for Computational Linguistics.

Lifeng Han. 2018. Machine translation evaluation resources and methods: a survey. In *IPRC – Irish Postgraduate Research Conference*, Dublin, Ireland.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Dominik Macháček and Ondřej Bojar. 2020. Presenting simultaneous translation in limited space. In *Proceedings of the 20th Conference Information Technologies – Applications and Theory (ITAT 2020), Hotel Tyrapol, Oravská Lesná, Slovakia, September 18-22, 2020*, volume 2718 of *CEUR Workshop Proceedings*, pages 34–39. CEUR-WS.org.

Markus Müller, Sarah Fünfer, Sebastian Stüker, and Alex Waibel. 2016a. Evaluation of the KIT lecture translation system. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1856–1861, Portorož, Slovenia. European Language Resources Association (ELRA).

Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. 2016b. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 82–86, San Diego, California. Association for Computational Linguistics.

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2016. Dynamic transcription for low-latency speech transla-

tion. In *17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016*, volume 08-12-September-2016 of *Proceedings of the Annual Conference of the International Speech Communication Association. Ed. : N. Morgan*, pages 2513–2517. International Speech and Communication Association, Baixas.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-latency neural speech translation. In *Interspeech 2018*, Hyderabad, India.

Ofir Press and Noah A. Smith. 2018. You may not need attention. *CoRR*, abs/1810.13409.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Hao Xiong, Ruiqing Zhang, Chuanqiang Zhang, Zhongjun Hea, Hua Wu, and Haifeng Wang. 2019. Dutongchuan: Context-aware translation model for simultaneous interpreting.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. Simpler and faster learning of adaptive policies for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1349–1354, Hong Kong, China. Association for Computational Linguistics.