Accuracy of AI-generated Captions With Collaborative Manual Corrections in Real-Time

Korbinian Kuhn
kuhnko@hdm-stuttgart.de
Stuttgart Media University
Stuttgart, Germany

Verena Kersken
kersken@hdm-stuttgart.de
Stuttgart Media University
Stuttgart, Germany

Gottfried Zimmermann
zimmermanng@hdm-stuttgart.de
Stuttgart Media University
Stuttgart, Germany

ABSTRACT

Automatic Speech Recognition (ASR) is a cost-efficient and scalable tool to automate real-time captioning. Even though its overall quality has improved rapidly, generated transcripts can be inaccurate. While manual correction helps to increase transcription accuracy, this causes new real-time challenges, especially for live-streaming. Crowd-sourcing can make the high workload more manageable by distributing the work across multiple individuals. In this paper, we developed a prototype that enables humans to collaboratively correct AI-generated captions in real-time. We conducted an experiment with 40 participants to measure the accuracy of the created and manually corrected captions. The results show that manual corrections improved the overall text accuracy according to multiple metrics as well as overall qualitative analysis.

CCS CONCEPTS

• **Human-centered computing → Accessibility technologies**; *Empirical studies in accessibility*; *Accessibility design and evaluation methods.*

KEYWORDS

captioning; real-time; automatic speech recognition; crowd-sourcing; subtitles;

ACM Reference Format:
Korbinian Kuhn, Verena Kersken, and Gottfried Zimmermann. 2023. Accuracy of AI-generated Captions With Collaborative Manual Corrections in Real-Time. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3585724

1 INTRODUCTION

Captions are a necessary accessibility tool for deaf and hard-of-hearing (DHH) individuals. For pre-recorded media, captions are created after the recording, and displayed on the screen during playback. Captions for live-events are much harder to create and thus can limit individual participation. In Higher Education settings, online lectures, talks and seminars are increasingly common and need to be accessible to all students. Institutions therefore need a scalable and cost-efficient solution to provide accurate real-time captions.

Today, there are two common solutions that enable real-time transcription: Captions can either be generated manually by professionals using stenography or respeaking, or automatically through Automatic Speech Recognition (ASR). Neither of the current approaches fulfill the requirements of Higher Education institutions. Professionals are costly and scarce, particularly for specialist subject areas. ASR, on the other hand, is cheap and widely available, but often lacks accuracy to the degree that the resulting transcript may not be understandable to readers.

One promising solution to this problem is a semi-automated workflow, in which non-professionals (in terms of the provision of captioning) correct automatically generated captions. Correction has the advantage that humans can focus on identifying and editing significant errors, rather than trying to type every spoken word. A crowd-based approach would further reduce the individual workload. Evidence suggests that non-professionals can correct ASR-generated transcripts and significantly improve the accuracy [8, 25, 31].

This semi-automated workflow is particularly promising for Higher Education, as many lectures take place (simultaneously), solutions cannot be overly costly, and students generally have advanced computer skills. A possible scenario is, that tutors or other students attending the lecture could correct ASR-generated captions in real-time. DHH-students could then view the corrected captions in the same software.

The accuracy of captions is often measured using the Word Error Rate (WER) which is a ratio of errors in the transcript to the total number of words spoken. This measure has, however, been challenged as it does not necessarily reflect the usefulness of captions [11, 15, 23, 24, 30, 35]. It remains an open question whether AI-generated captions, that are corrected by a group of people in real-time, are sufficiently accurate.

To assess this semi-automated workflow, we conducted a user test, in which Higher Education students corrected AI-generated captions for a video lecture. We therefore developed a prototype that enables collaborative corrections in real-time and measuring the editing in a controlled environment. We used multiple metrics and various approaches to evaluate caption accuracy. In summary our research evaluates how collaborative real-time editing, supported by a specialised application interface, affects the accuracy of captions.

2 BACKGROUND

Captions greatly improve the accessibility of videos and online events for people who are deaf and hard-of-hearing (DHH). One

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585724

context in which captioning is particularly relevant is Higher Education. Remote lectures, online meetings and digital events are increasingly part of university studies and need to be made accessible to all students. Captions do not only benefit DHH students [39], but also support second language learners [37], make non-native instructors easier to understand [17], and help students acquire specialist vocabulary [7]. Furthermore, captions make video content searchable [13].

2.1 Creating captions in real-time

In Higher Education settings, captioning for live events is provided through professionals using either stenography or respeaking combined with manual correction. These professionals work in teams of two or more for longer events. This process is not only laborious and costly; there is also a shortage of professionals to provide these services [34].

Additionally, respeakers create non-verbatim transcripts as they lag around 20 to 40 words per minute (wpm) behind the original speakers. Edited captions are favored by academics and many viewers, as they are easier to read and compress information. Many deaf associations, on the other hand, consider editing as censorship, which is critical, as respeaking cannot produce a verbatim transcript [29]. This also raises the question of whether qualitative measures, such as an evaluation by DHH individuals, are necessary to make judgements about the quality and accuracy of captions.

Two alternatives to captioning by professional transcribers are manual transcription by non-professionals (in terms of providing captioning services) or automatic speech recognition.

Even for advanced typers, it is difficult to provide a verbatim transcript, as speaking is up to ten times faster than writing [34]. With crowd-sourcing, the task of transcription can be divided, and each person only has to process fractions of the whole text. Lasecki et al. [18] asked groups of people to manually create captions for a talk and developed an algorithm to merge participants' individual keyboard input into a single text stream. Their results suggest that this method of crowd-sourcing captions can outperform individuals, ASR and professionals in terms of coverage and latency.

The other alternative to professional transcription is ASR. With the increasing computing power of artificial intelligence (AI), ASR is now part of many common software packages, and can generate captions in real-time. Commercial solutions use phrases like "state-of-the-art accuracy" [14], "produce accurate transcripts" [3] or "high quality transcription" [21]. In a comparison of different industry leaders in 2022, their services reached accuracy rates from 69% to 88% for high-resource languages like English and German [4]. In terms of research various deep-learning models target benchmarks like the common LibriSpeech ASR corpus [26]. From 2015 to 2021, the most accurate models could decrease the WER from 13.25% to 2.5% [38]. However, the lowest error rates are currently only achieved by state-of-the-art end-to-end models, while a large proportion of commercial ASR systems are still based on hybrid systems which are composed of separate acoustic, language, and pronunciation models [20]. It is therefore important to distinguish between a high accuracy achieved with specific data sets and a general usage in real world applications.

The accuracy of ASR is highly dependent on audio quality, speaker and terminology. Some sources of error are background noise [16], the speakers' gender and accent [5, 12], and the use of specialist vocabulary or abbreviations [2]. Another major influence on ASR quality is the language itself. For languages other than English, transcripts tend to be less accurate, presumably because of smaller training sets and additional sources of error, such as capitalization or cases.

Despite the promising results of ASR, Deshpande et al. [10] note that the accuracy of current speech recognition tools is insufficient for live university lectures that comprise specialist vocabulary. The authors raise the issue that even with improvement of ASR technology, some limitations regarding technical terms or challenging aspects of conversational speech are likely to remain. Analyzing AI-generated captions for lecture material, Parton [27] comes to a similar assessment. Without editing, the quality of automatically generated captions is insufficient for universities to meet their legal obligation to provide accessible learning material.

2.2 Semi-automated workflow

An alternative to automatic or manual transcription, is a semi-automated workflow described by Wald [33] that uses ASR to generate captions and humans to correct errors. Che et al. [8] report that using ASR-generated captions as draft for manual caption production reduces the error rate of the captions by about half (54.3%). Furthermore, the working time of professionals can be shortened by 54% compared to manual caption creation. Their results align with a test conducted by Soe et al. [31] that also found that manual correction decreases the WER of ASR-generated captions. In a similar user study, Munteanu et al. [25] also see a relative WER reduction of 53%.

Manual correction might also target errors that influence meaning, rather than minor grammatical errors that do not impact the understanding of the content [6, 17]. Deshpande et al. [10] evaluated collaborative editing in a non-live scenario and gave groups of students the task to correct ASR-generated captions for lecture recordings over a period of five days. Afterwards, DHH users were asked to judge the usefulness of these captions and they rated them as very accurate – and consistently rated them higher than captions generated by professional transcription services.

Multiple studies suggest further investigation into collaborative editing of ASR-generated captions [10, 17, 33] and also consider metrics other than WER for evaluation [17]. A previous study found that students were able to increase the accuracy of automatically generated captions in a real-time scenario [28]. They found that the participants performed better, when captions were presented simultaneously to the spoken words, as compared to captions presented by a 1-second delay to the audio. However, their study uses a Wizard of Oz paradigm to simulate the editor, and resulting usability issues reportedly impacted participants' performance.

2.3 Metrics to measure transcription accuracy

The most common metric to measure transcription accuracy is the Word Error Rate (WER). While it is a practical metric to compare different ASR systems, multiple studies criticize that it not necessarily reflects the understandability of a text. It penalizes all types of errors equally and therefore does not always align with human perception of accuracy, as humans judge errors distorting the meaning more harshly [23]. It can also be misleading regarding spoken language understanding [35] and not adequately reflect errors of low-frequency yet important words [2]. Favre et al. [11] state that rather than pure text accuracy, ASR output should be validated for real-world use-cases and therefore how useful it is for humans.

Morris et al. [24] claim that Match Error Rate (MER) performs better than WER regarding the proportion of information communicated. They also present Word Information Lost (WIL) as a simple approximation to the proportion of word information lost. However, their studies focus on Connected Speech Recognition applications, which have different needs than captioning software. Another measure is the Character Error Rate (CER), that is calculated like WER but on a character basis. It is more robust against minor typing mistakes, e.g. resulting from manual keyboard corrections.

Romero-Fresco and Pérez [30] developed a metric to measure the accuracy of respeaking and automatic captioning. The NER-model is based on the total number of words in a respoken text, editing errors caused by the respeaker and recognition errors resulting from mispronunciations or the ASR system. The errors are differently weighted for calculation and therefore must be manually classified as serious, standard or minor errors. They further state that live subtitles may be expected to reach an accuracy of 98%. Even though the NER-model is suited for a respeaking workflow, it might also be used to evaluate manually corrected ASR transcripts.

Kafle and Huenerfauth [15] introduce the Automated-Caption Evaluation (ACE) metric, that uses a machine learning approach to particularly measure the accuracy of captions. Their study shows that ACE correlates better with their DHH participants’ subjective scores on the usability of captions than WER. Conversely, Wells et al. [36] did not measure a statistically relevant difference between WER and ACE for live captioning.

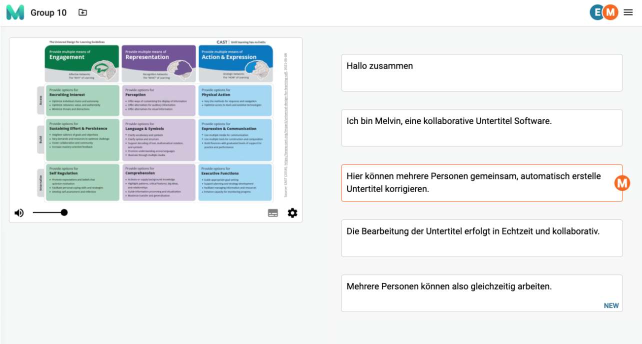
Most research and commercial solutions use WER to measure the accuracy of captions. At the same time to the best of our knowledge there is no WER threshold at which a transcript can be seen as “sufficiently accurate for accessibility purposes”. While Microsoft considers a WER of 5%-10% to be good, and a WER of 20% acceptable [22], standards like Web Content Accessibility Guidelines (WCAG) [1] or Federal Communications Commission (FCC) [9] do not specify any requirements for the accuracy of accessible captions. The W3C Web Accessibility Initiative (WAI) even defines automatic captions as not sufficient, unless they are confirmed to be fully accurate/error-free, and state that usually significant editing is needed [19].

3 METHOD

3.1 Participants

40 people participated in the study. 35 were undergraduate Computer Science or Education students; five were research staff at multiple universities. Participation was voluntary. Prior to taking part, participants were informed about the goals of the study, what participation entailed, and gave their informed consent to take part.

Participants edited the captions in groups of three or groups of five. We assume these group sizes to be practical in medium-sized courses and that the parallel editing is manageable without further coordination. Participants themselves chose a time slot for their participation. The number of people in a group was assigned according to participants’ availability.


Group 10
E
M
Provide multiple means of Engagement
Provide multiple means of Representation
Provide multiple means of Action & Expression
Recruiting Interest
Perception
Physical Action
Sustaining Effort & Persistence
Language & Symbols
Expression & Communication
Self Regulation
Comprehension
Executive Functions
M
NEW

Hallo zusammen

Ich bin Melvin, eine kollaborative Untertitel Software.

Hier können mehrere Personen gemeinsam, automatisch erstelle Untertitel korrigieren.

Die Bearbeitung der Untertitel erfolgt in Echtzeit und kollaborativ.

Mehrere Personen können also gleichzeitig arbeiten.

Figure 1: Interface of prototype

3.2 Video

Participants were asked to edit captions for a 12-minute video recording, containing 1363 words in total and an average of 110 wpm. The language used was German, with some technical terms in English. The video was a pre-recorded conference talk on the topic of “Universal Design for Learning and Digital Accessibility”. Although it is a limited representation of content for Higher Education, it can be used to evaluate the general workflow of real-time editing. The topic of the video was broadly connected to the participants’ field of study, but it is unlikely that they were familiar with the specific content. The video was recorded with a PC in non-professional conditions with average audio quality. A Jabra speakerphone was used as microphone for the recording. The male speaker spoke in a natural, conversational manner, which included hesitation vowels, unclear pronunciation, incomplete sentences, and grammatical errors.

3.3 Captions

ASR-generated captions for the video were produced prior to the user test. The recording was delivered to the ASR engine as a complete file at once, not as an audio stream. We did not add a custom vocabulary to improve ASR results regarding specialist vocabulary or abbreviations. Captions for the user test were generated by Microsoft Azure, as it delivered the most accurate result for the recording compared to other providers (Google, Amazon, IBM, Panopto and Amberscript). The resulting transcript was not corrected but split manually into chunks according to WCAG 2.2 [1] guidelines for captions and extended with timestamps. The ASR-generated captions were presented simultaneously to the spoken words in the video to reduce the cognitive load for the participants [28].

3.4 Interface

The editor is shown in Fig. 1. The interface follows the design of existing solutions of caption editors. The video is displayed on the left without any controls to start, stop or rewind it. The captions appear consecutively on the right below the existing content. Each caption is displayed as an input field. To avoid conflicts, only one user could edit a caption at a time. Additionally, captions that are currently edited are highlighted in a user specific color, to give a visual indication to the other participants.

The editors' frontend is a web application that works in all major browsers. To enable real-time collaboration, the state is synchronized with a backend providing a REST API and a WebSocket API. All corrections made by users, or the insertion of new captions by the ASR-system appear without a page reload. Therefore participants can keep track of the group work without losing focus on currently selected elements, e.g. an input field. If a user is not editing a caption, the application is automatically scrolling to the newest captions.

3.5 Procedure

User tests were conducted remotely and in-person. For in-person tests, groups were invited to a university computer lab. Participants wore headphones and worked on their own laptops during the user test. For remote user tests, participants joined a video call (Big Blue Button) from their respective location.

First, participants were informed about the goals of the study and what participation entailed. After giving their informed consent to take part, each participant received an anonymous user-id to join the collaborative editor (see Fig. 1) via their browser. Before the test started, participants were instructed to correct the captions accompanying the video. They did not receive any specific instructions as to which errors they should or should not correct or that quality or speed mattered. Participants could edit captions by clicking on the area displaying the caption segment. This caption would then be locked so that other users could not edit.

The video was shown at normal speed and not paused or stopped during the editing process. New captions appeared in synchronization with the videos' audio track. While the video was playing, participants could not interact with each other via chat or other means.

After the video ended, participants were asked to leave the application and to answer a questionnaire that contained questions about their subjective experience of the task, such as perceived stress level, and whether they could retain the information presented in the video. Participants were debriefed and received a small thank-you gift for their participation.

3.6 Analysis

3.6.1 Error metrics. We used the JiWER [32] python implementation to calculate error rates based on the Levenshtein minimum-edit distance between a ground-truth and a hypothesis sentence. Besides the most common Word Error Rate (WER), we also measured the Character Error Rate (CER), the Match Error Rate (MER) and Word Information Lost (WIL). If not stated otherwise, the error rates were calculated with the following text transformations: lowercase (case-insensitivity), removed punctuation and removed duplicate spaces. According to German grammar rules, all numbers from one to twelve were replaced by their written word.

3.6.2 Caption correction state categorization. We classified each caption by its correction state based on text equality compared to the ASR and the ground-truth transcription. We differentiated between five states: (1) Captions with errors that were fully corrected and (2) captions without errors that stayed unmodified were categorized as correct. (3) Captions that were missing corrections or (4) contained faulty modifications were categorized as faulty. (5) Captions that were partly corrected are classified as an unknown state. As the comparison was done on a segmentation basis, moving words to another segment, the transcript would still be correct but result in an error for this classification.

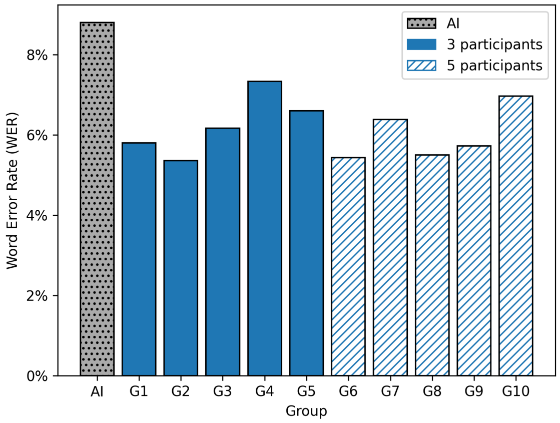


Figure 2: Word Error Rate of AI and group corrections

3.6.3 Classification of serious errors. In order to explore a possible qualitative measure of captions, we asked three people (one DHH individual) to read the ASR-generated transcript and highlight any errors that they felt compromised the meaning of the text. The rating aims to reflect the perspective of people who depend on captions to understand the text and cannot rely on additional information from the audio track. Therefore raters were just given the transcript and did not watch the video prior to marking errors. Raters may have missed some mistakes, as the intended meaning of the text was not always clear to them. We then automatically searched the transcripts for the reported errors and whether they were modified or corrected. In contrast to the automated metrics, this qualitative measure focuses on individual's subjective assessment of whether the content of the text can be understood and its readability.

4 RESULTS

4.1 Word Error Rate of AI and group corrections

Manual corrections did improve the overall text accuracy of the AI-generated captions. The initial WER of the AI (8.8%) was reduced to a mean WER of 6.1% (-2.7 pp). Across all participating groups, the WER decreased by 1.5 to 3.4 percentage points, resulting in WERs of 5.4 to 7.3% (see Fig. 2). A Mann-Whitney-U-Test showed that there was no significant difference (U = 10, p = 0.690) between the two different group sizes. The groups of three have a median WER of 6.2% and the groups of five a median WER of 5.7%.

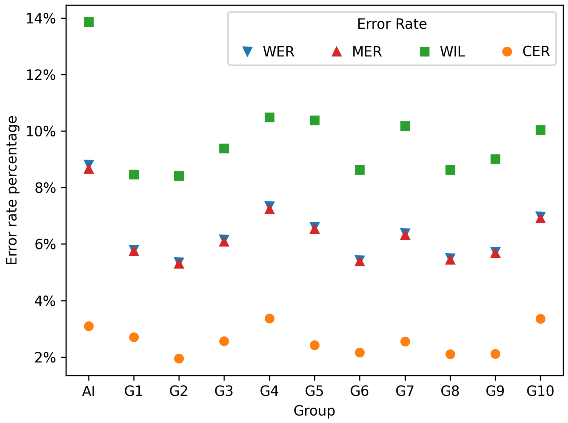


Figure 3: Comparison of error metrics

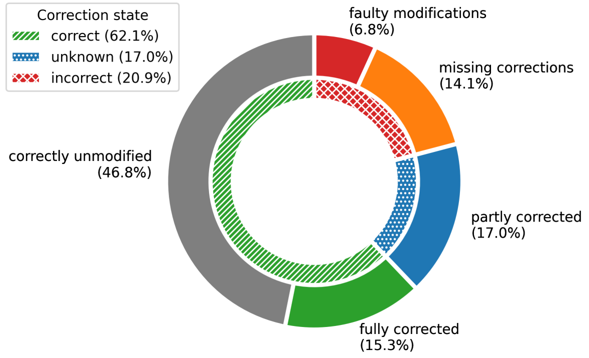


Figure 4: Caption segments categorized by correction state

4.2 Comparison of error metrics

The initial WER of the AI (8.8%) was reduced across all groups to a minimum of 5.4%, a maximum of 7.3% and to a mean of 6.1% (see Fig. 3). The initial MER of the AI (8.7%) was reduced across all groups to a minimum of 5.3%, a maximum of 7.2% and to a mean of 6.1%. The initial WIL of the AI (13.9%) was reduced across all groups to a minimum of 8.4%, a maximum of 10.5% and to a mean of 9.4%. The initial CER of the AI (3.1%) was reduced across all groups to a minimum of 2.0%, a maximum of 3.4% and to a mean of 2.5%.

4.3 Caption count by correction state

Of 123 caption segments in total, 57 (46,3%) required modifications of the AI text to match the ground-truth and 66 (53,7%) were initially correct. Fig. 4 shows that on average, 46.8% were left correctly unmodified and 15.3% were fully corrected, resulting in 62.1% captions categorized as correct. 17.0% were partly corrected and categorized as an unknown correction state. 14.1% were missing modifications and 6.8% contained faulty modifications, resulting in 20.9% caption segments categorized as incorrect.

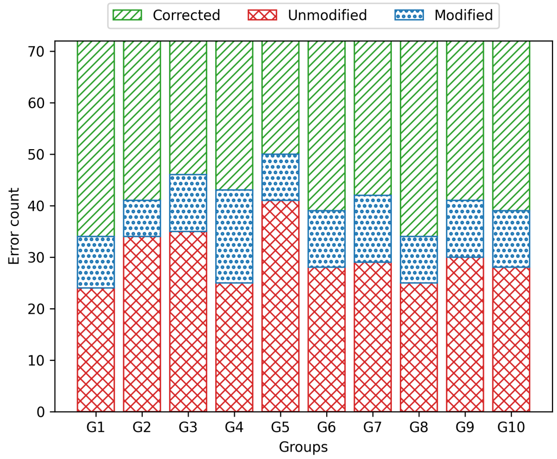


Figure 5: Corrections of serious errors

4.4 Corrections of serious errors

In total, 72 errors distributed over 53 caption segments were classified as serious. Of these, 22 to 38 (mean=31.1) were corrected and 24 to 41 (mean=29.9) were left unmodified (see Fig. 5). Between 7 to 18 (mean=11.0) were modified, but did not result in a correct state (e.g. were partly or faulty corrected). Of 38 captions that contain exactly one error, on average 19.2 were fully corrected, 5.5 modified and 13.3 left unmodified. Of 15 captions that contain more than one error, on average 1.3 were fully corrected, 1.9 partly corrected, 4.4 modified and 7.4 left unmodified.

4.5 Word Error Rate with different text transformations

Table 1 shows the calculated WER based on different text transformations. The initial WER of the AI is between 16.4% without text transformations and 8.8% with all transformations combined. The WER decreased by the group corrections regarding the transformation level between 2.7 and 3.4 percentage points. The AI transcript required between 120 and 224 modifications to match the reference solution. For the corrected transcripts, the average modifications needed to match the reference solution are between 83.5 and 177.5.

5 DISCUSSION

Results of the user test show that the manual correction of ASR-generated captions by non-professionals reduced the overall WER in the final text. Similar findings have been reported before, here we show that they also apply for the correction of captions in real-time. People working together in groups of five did not correct significantly more errors than those working in groups of three. The ASR-generated transcript had an initial WER of 8.8% and manual correction reduced the WER of the transcript to 6.1% on average. Related studies reported a WER reduction of around 50% [8, 25]. In our study, the decrease was only around 30%, which might result

Table 1: Word Error Rate with different text transformations

	Word Error Rate			Required modifications		
Text transformation	AI	Groups	Decrease	AI	Groups	Decrease
None	16.4%	13.0%	3.4 pp	224	177.5	46.5
No Punctuation	11.4%	8.5%	2.9 pp	155	115.9	39.1
Lowercase	14.6%	11.3%	3.3 pp	199	154.4	44.6
Lowercase + No Punctuation	8.8%	6.1%	2.7 pp	120	83.5	36.5

from the fact that the editing had to happen in real-time. Since there is no defined threshold that marks captions as sufficiently accurate to be accessible, and considering the amount of remaining errors in the edited captions, it is unclear whether the corrected captions are accessible to DHH individuals.

All metrics (WER, CER, MER, WIL) measuring the captions' accuracy show similar results. It therefore appears that WER and these alternative measures can be used interchangeably, at least for texts with a relatively low overall error rate. These metrics are heavily dependent on pre-processing steps that transform the text before calculation of the edit distance. The average WER for unmodified texts is almost twice as high compared to a calculation that ignores punctuation and capitalization. However, especially for a language like German, these transformations have a big impact on text understanding and readability. Therefore, error rates calculated with heavy pre-processing can be misleading.

Human readers classified 72 errors as serious with regard to text understanding in the ASR-generated transcript. On average, 43% of these errors were corrected through manual editing. This is a slightly higher decrease compared to the WER reduction of 30%. Therefore, manual corrections seem to have a bigger impact on meaningful errors, than pure text accuracy. This supports the findings of [6, 17], who state that manual correction might target errors that influence meaning, rather than minor grammatical mistakes. While 50% of caption segments with a single error were fully corrected, only 20% with multiple errors were at least partly corrected. This could be due to several reasons. It may be that correcting multiple errors is more challenging, because people do not perceive all errors. Alternatively, the editor's interface and the blocking of text could make multiple errors harder to spot.

Evaluating the accuracy on a per segment basis, the results show that 39% of the segments were modified, but only 15% of them resulted in a fully correct state. 7% were initially correct but were made incorrect by users' modifications. In some cases manual correction therefore introduced additional errors into the transcript. These faulty modifications are irrelevant if they result from moving words to the previous or next caption. Text optimizations like the removal of duplicate words or resulting from the speaker correcting themselves can also be seen as trivial, or even support the text's understandability. But it is also possible that people introduce errors because they simply misheard words, or try to fix mistakes based on their own contextual understanding of the text. The creation of new errors by manual correction is an interesting finding and requires further research.

6 CONCLUSION

A semi-automated workflow combining ASR and crowd-sourced manual corrections seems promising for real-time events because people help reduce the overall number of errors in the text. The result is a verbatim transcript, in contrast to transcripts created by respeakers, who reduce the overall number of words. Our analysis indicates that the accuracy is not satisfactory and needs to be addressed by further optimizations of the workflow and user-interface.

Common quantitative measures of accuracy showed similar results and seem to be used interchangeably for texts with low overall error rates. It remains an open question if metrics like WER are only a measurement for text accuracy or whether a threshold can classify captions as sufficiently accurate for DHH individuals. Until then, alternative qualitative measures which reflect meaningful errors and text understandability are required.

There is a need in Higher Education for scalable and cost-efficient solutions to provide accurate real-time captions. For now, ASR alone is not a sufficient solution for DHH students. Despite the limited scope of this study, the developed prototype could effectively simulate a semi-automated workflow and the results show that extending ASR with manual corrections seems to be a promising approach.

7 FUTURE WORK

The interface might be limited by the approach of using segment-based editing. The accuracy of captions could increase if people correct more meaningful chunks like complete sentences or work together on a continuous text. Furthermore, the ideal group size and how the user interface can support collaboration should be examined. An open question is whether students can correct captions while also paying attention to a lecture, or whether additional people e.g. tutors are required. Comparative measurements of cognitive load and subjective impressions are needed to improve the user experience. Besides workflow and interface optimizations, more accurate measures are required, to classify when captions are sufficiently accurate to count as accessible. They need to be scalable and easy to adapt, like a qualitative error classification but without human involvement.

ACKNOWLEDGMENTS

We thank our participants for their time and contribution to this research. We thank Benedikt Reuter and Niklas Egger for developing the prototype. This work was conducted as part of the SHUFFLE Project ("Hochschulinitiative Barrierefreiheit für alle") and funded by "Stiftung Innovation in der Hochschullehre".

REFERENCES

[1] Chuck Adams, Alastair Campbell, Michael Cooper, and Andrew Kirkpatrick. 2021. *Web Content Accessibility Guidelines (WCAG) 2.2.* World Wide Web Consortium (W3C). Retrieved January 12, 2023 from https://www.w3.org/TR/WCAG22/

[2] Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How Might We Create Better Benchmarks for Speech Recognition?. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future.* Association for Computational Linguistics, Online, 22–34. https://doi.org/10.18653/v1/2021.bppf-1.4

[3] Amazon. 2023. AWS - Amazon Transcribe Features. Retrieved January 18, 2023 from https://aws.amazon.com/transcribe/features/

[4] Sheryl Ballenger. 2022. Access for Deaf and Hard of Hearing Individuals in Informational and Educational Remote Sessions. *Assistive Technology Outcomes & Benefits* 16, 2 (2022), 45–55.

[5] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, Richard Rose, Vivek Tyagi, and Christian Wellekens. 2007. Automatic Speech Recognition and Speech Variability: A Review. *Speech Commun.* 49, 10–11 (oct 2007), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

[6] Bhavya Bhavya, Si Chen, Zhilin Zhang, Wenting Li, Chengxiang Zhai, Lawrence Angrave, and Yun Huang. 2022. Exploring collaborative caption editing to augment video-based learning. *Educational technology research and development* 70, 5 (01 Oct 2022), 1755–1779. https://doi.org/10.1007/s11423-022-10137-5

[7] Stephen Bird and John Williams. 2002. The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling. *Applied Psycholinguistics* 23, 4 (2002), 509–533. https://doi.org/10.1017/S0142716402004022

[8] Xiaoyin Che, Sheng Luo, Haojin Yang, and Christoph Meinel. 2017. Automatic Lecture Subtitle Generation and How It Helps. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT).* IEEE Computer Society, Timisoara, Romania, 34–38. https://doi.org/10.1109/ICALT.2017.11

[9] Federal Communications Commission. 2021. Closed Captioning on Television. Retrieved January 18, 2023 from https://www.fcc.gov/consumers/guides/closed-captioning-television

[10] Rucha Deshpande, Tayfun Tuna, Jaspal Subhlok, and Lecia Barker. 2014. A crowdsourcing caption editor for educational videos. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings.* IEEE Computer Society, Madrid, Spain, 1–8. https://doi.org/10.1109/FIE.2014.7044040

[11] Benoit Favre, Kyla Cheung, Siavash Kazemian, Adam Lee, Yang Liu, Cosmin Munteanu, Ani Nenkova, Dennis Ochei, Gerald Penn, Stephen Tratz, Clare Voss, and Frauke Zeller. 2013. Automatic Human Utility Evaluation of ASR Systems: Does WER Really Predict Performance?. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013.* ISCA, Lyon, France, 3463–3467.

[12] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. https://doi.org/10.48550/ARXIV.2103.15122

[13] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. 2001. Audio Partitioning and Transcription for Broadcast Data Indexation. *Multimedia Tools and Applications* 14, 2 (01 Jun 2001), 187–200. https://doi.org/10.1023/A:1011303401042

[14] Google. 2023. Google Cloud - Speech-to-Text. Retrieved January 18, 2023 from https://cloud.google.com/speech-to-text/

[15] Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the Usability of Automatically Generated Captions for People Who Are Deaf or Hard of Hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17).* Association for Computing Machinery, New York, NY, USA, 165–174. https://doi.org/10.1145/3132525.3132542

[16] Keisuke Kinoshita, Tsubasa Ochiai, Marc Delcroix, and Tomohiro Nakatani. 2020. Improving Noise Robust Automatic Speech Recognition with Single-Channel Time-Domain Enhancement Network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, Barcelona, Spain, 7009–7013. https://doi.org/10.1109/ICASSP40776.2020.9053266

[17] Raja Kushalnagar, Walter Lasecki, and Jeffrey Bigham. 2012. A Readability Evaluation of Real-Time Crowd Captions in the Classroom. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (Boulder, Colorado, USA) *(ASSETS '12).* Association for Computing Machinery, New York, NY, USA, 71–78. https://doi.org/10.1145/2384916.2384930

[18] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-Time Captioning by Groups of Non-Experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12).* Association for Computing Machinery, New York, NY, USA, 23–34. https://doi.org/10.1145/2380116.2380122

[19] Shawn Lawton Henry, Geoff Freed, and Judy Brewer. 2022. *Making Audio and Video Media Accessible - Captions/Subtitles.* Web Accessibility Initiative. Retrieved January 12, 2023 from https://www.w3.org/WAI/media/av/captions/

[20] Jinyu Li. 2022. Recent advances in end-to-end automatic speech recognition.

[21] Microsoft. 2023. Azure - Speech to Text. Retrieved January 18, 2023 from https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/

[22] Microsoft. 2023. Microsoft - Test accuracy of a Custom Speech model. Retrieved January 18, 2023 from https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data

[23] Taniya Mishra, Andrej Ljolje, and Mazin Gilbert. 2011. Predicting Human Perceived Accuracy of ASR Systems.. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011.* ISCA, Florence, Italy, 1945–1948. https://doi.org/10.21437/Interspeech.2011-364

[24] Andrew Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing.* ISCA, Jeju Island, Korea, 2765–2768. https://doi.org/10.21437/Interspeech.2004-668

[25] Cosmin Munteanu, Ron Baecker, and Gerald Penn. 2008. Collaborative Editing for Improved Usefulness and Usability of Transcript-Enhanced Webcasts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08).* Association for Computing Machinery, New York, NY, USA, 373–382. https://doi.org/10.1145/1357054.1357117

[26] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, South Brisbane, QLD, Australia, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

[27] Becky Parton. 2016. Video Captions for Online Courses: Do YouTube's Auto-generated Captions Meet Deaf Students' Needs? *Journal of Open, Flexible, and Distance Learning* 20, 1 (August 2016), 8–18. https://www.learntechlib.org/p/174235

[28] Patricia Piskorek, Nadine Sienel, Korbinian Kuhn, Verena Kersken, and Gottfried Zimmermann. 2022. Evaluating collaborative editing of ai-generated live subtitles by non-professionals in German university lectures. In *Assistive Technology, Accessibility and (e)Inclusion: 18th International Conference, ICCHP-AAATE 2022, Lecco, Italy, July 11–15, 2022, Open Access Compendium, Part I.* ICCHP, Lecco, Italy, 165–175.

[29] Pablo Romero-Fresco. 2009. More haste less speed: Edited versus verbatim respoken subtitles. *Vigo International Journal of Applied Linguistics* 6 (01 2009), 109–133.

[30] Pablo Romero-Fresco and Juan Martínez Pérez. 2015. *Accuracy Rate in Live Subtitling: The NER Model.* Palgrave Macmillan UK, London, UK, 28–50. https://doi.org/10.1057/9781137552891_3

[31] Than Htut Soe, Frode Guribye, and Marija Slavkovik. 2021. Evaluating AI Assisted Subtitling. In *ACM International Conference on Interactive Media Experiences* (Virtual Event, USA) *(IMX '21).* Association for Computing Machinery, New York, NY, USA, 96–107. https://doi.org/10.1145/3452918.3458792

[32] Nik Vaessen. 2022. JiWER: Similarity measures for automatic speech recognition evaluation. Retrieved November 11, 2022 from https://pypi.org/project/jiwer/

[33] Mike Wald. 2006. Captioning for Deaf and Hard of Hearing People by Editing Automatic Speech Recognition in Real Time. In *Proceedings of the 10th International Conference on Computers Helping People with Special Needs* (Linz, Austria) *(ICCHP'06).* Springer-Verlag, Berlin, Heidelberg, 683–690. https://doi.org/10.1007/11788713_100

[34] Mike Wald. 2006. Creating Accessible Educational Multimedia through Editing Automatic Speech Recognition Captioning in Real Time. *Interactive Technology and Smart Education* 3 (05 2006), 131–141. https://doi.org/10.1108/17415650680000058

[35] Ye-Yi Wang, A. Acero, and C. Chelba. 2003. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721).* IEEE, St Thomas, VI, USA, 577–582. https://doi.org/10.1109/ASRU.2003.1318504

[36] Tian Wells, Dylan Christoffels, Christian Vogler, and Raja Kushalnagar. 2022. Comparing the Accuracy of ACE and WER Caption Metrics When Applied to Live Television Captioning. In *Computers Helping People with Special Needs: 18th International Conference, ICCHP-AAATE 2022, Lecco, Italy, July 11–15, 2022, Proceedings, Part I* (Lecco, Italy). Springer-Verlag, Berlin, Heidelberg, 522–528. https://doi.org/10.1007/978-3-031-08648-9_61

[37] Paula Winke, Susan Gass, and Tetyana Sydorenko. 2010. The Effects of Captioning Videos Used for Foreign Language Listening Activities. *Language Learning and Technology* 14 (02 2010), 65–86.

[38] Papers with Code. 2023. Speech Recognition on LibriSpeech test-other. Retrieved January 18, 2023 from https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-other

[39] Joong-O Yoon and Minjeong Kim. 2011. The Effects of Captions on Deaf Students' Content Comprehension, Cognitive Load, and Motivation in Online Learning. *American annals of the deaf* 156 (06 2011), 283–9. https://doi.org/10.1353/aad.2011.0026