Caption This! Best practices for live captioning of jargon-rich scientific presentations

Michele L. Cooke[1], Celia R. Child[2], Elizabeth C. Sibert[3], Christoph von Hagke[4] and S. G. Zihms[5]

Affiliations:
[1]University of Massachusetts Amherst, Department of Geosciences.
[2]Bryn Mawr College, Department of Geology.
[3]Harvard University, Department of Earth and Planetary Sciences.
[4]University of Salzburg, Department of Geography and Geology.
[5]University of Western Scotland, UWS Academy.
*Correspondence to: cooke@umass.edu

Whether your scientific presentation is in-person or remote, everyone will understand more of your presentation if it has captions. Like subtitles of a movie, open captioning makes verbal material accessible for many people. A study of BBC television watchers reports that 80% of caption users are not deaf nor hard of hearing (Ofcom, 2006). During English-spoken scientific presentations, people who are deaf or hard of hearing, people who have auditory processing disorder and not yet fluent non-native English speakers develop listening fatigue that can prohibit their understanding and limit their participation in discussions. Increasing the accessibility of our presentations and improving inclusivity of discussions provides a path towards increasing diversity within sciences. Studies show that subtitles/captioning improve both English language skills (e.g., Vanderplank, 2016; Wang & Liu, 2011) and accessibility of science for deaf and hard of hearing participants (e.g., Kawas et al., 2016; Vanderplank, 2016). Furthermore, not everyone may be in a space where they can access audio, for example, if they are sharing space with other workers.

A myriad of tools and platforms can provide captioning for live presentations. Why then don't we regularly caption presentations? Our resistance may be due to factors such as not knowing or believing that captioning is needed, not knowing how to use these tools, and believing that the resulting captioning will be inadequate. In response to the first reason, folks should not be forced to disclose their disability in order for presentations to be accessible to them. In response to the last two reasons, this article outlines different strategies for providing captions and presents results of our performance assessment of Artificial Intelligence (AI) based auto-caption of jargon rich geologic passages. Because most scientific presentations are delivered using either Microsoft PowerPoint or Google Slides presentation software, we focus our performance assessment on the auto-captioning provided by these platforms. While a variety of tools can add captions to recorded lectures that can be edited to improve accuracy, offering a transcript after a live presentation is not a suitable solution to improve participation. Here we provide evidence-based best-practices for providing captioning that will increase the accessibility of live scientific presentations

In-Person Presentations

For in-person presentations, trained human captionists or AI-based auto caption/transcription software can provide live captioning (Fig. 1). Captionists use stenography tools to provide

accurate transcription. In order for everyone to access the captions, the captionist's transcriptions can be projected onto a separate screen near the presentation slides.

Both *Microsoft Powerpoint* (with Office 365 or Presentation Translator) and *Google Slides* (with Chrome browser) provide built in AI-based auto-caption directly onto the presented slides that can be employed by anyone (instructions at Cooke & Caicedo, 20120). Third party software, such as *Ava*, *Rev*, *Otter.ai*, can also provide AI-based transcriptions. In addition to their availability, an advantage of *Slides/PowerPoint* over third-party transcription software is that the captioning is projected onto the same screen as the presentation. Having captions within the presenting slides frees the audience from having to shift their focus between presentation materials and a separate caption screen.

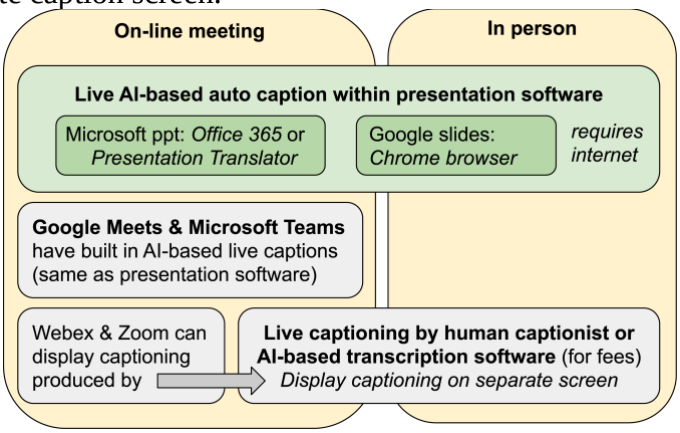

Fig. 1: Approaches and tools that can produce captioning of live scientific presentations. This study tests the performance of Microsoft Powerpoint and Google Slides AI-based auto-caption, which can be used for in-person and remote on-line presentations.

On-line presentations

For remote on-line presentations, any of the in-person strategies can also work. Human captionists anywhere in the world can join the remote meeting and provide captioning. In addition, the on-line meeting platforms *Google Meets* and *Micosoft Teams* offer built-in live auto-caption that use the same AI-based transcription tools as their presentation software. Within *Webex* and *Zoom*, captioning can be available to everyone if the host appoints the captionist within the meeting software. *Zoom* and *WebEx* also allow for third party auto-captions if the host has paid for those services. The benefits of providing captioning directly within *Microsoft PowerPoint* and *Google Slides* is that the AI-based captioning is built-in and you don't need to add another tool and pay for that service.

How accurate are captions for scientific talks?

If you have watched auto-captions provided by YouTube, then you have seen low quality captioning, sometimes called 'craptions' (Besner, 2019). The Word Error Rate (WER = incorrect words / total words) of *YouTube*'s non-ai-based auto-caption is 20-50%, which renders it useless unless creators manually edit the auto-generated transcript (Leduc, 2019). Typical word error types include split or blended words, incorrect spelling, incorrect guesses, etc. For both AI-based and human captioning, WER is impacted by microphone quality, internet quality, accent/style of the speaker and advance access of the captionist to the material. Jargon, such as often

encountered in science presentations, can be particularly challenging for accurate captioning. To challenge the performance of live auto-captioning software to capture scientific presentations, we chose two passages laden with geologic jargon (taken from Van der Pluijm & Marshak, 2004; Weil, 2006). Both passages have complex words that are not used outside of the discipline as well as common English words that are used differently by experts. For example, 'thrust' is typically a verb but geologists use it as an adjective for a type of fault. The second text also tests the recognition of acronyms. Prior to testing the auto-caption performance, we identified words that we expected to be challenging (Table 1).

Table 1: Words missed within captioning of American-accented English and standard sound quality.

Words that we expected the AI-based captions to miss	Words that that captions missed much of the time	Words that captions consistently missed
Nappes; substratum; lithosphere; vergent; accretionary; nonsubductable; radiogenic; Barrovian; metamorphism; paleomagnetic; Variscan; WEVB; Carboniferous; Permian; orocline; kinematic	Nappes; lithosphere; nonsubductable; Barrovian; Variscan; WEVB; orocline; granitic; phases; Blended words: 'thrusts and'; 'hinge zone'; 'WEVB's core'	Nappes; nonsubductable; Barrovian; Variscan*; WEVB*; orocline *captioned correctly under best practices and after some training

We measured the WER of *Microsoft Powerpoint* and *Google Slides* AI-based live auto captioning for both passages under a variety of conditions. WER indicates occurrence of error, so if the captioning never caught the acronym '*WEVB*', this would count as 4 mistakes in the second passage. With a recording of an American accented English female voice, we repeatedly tested the caption performance of both *PowerPoint* and *Slides*. For some tests, we decreased the sound quality by adding background noise and decreasing input volume. In another set of tests, we assess the WER of recordings of non-native English-speaking geologists reading the two passages. The accents (Chinese, Mexican, Spanish and German) are not meant to provide a complete accounting of the potential WER of non-native English speakers but instead show the relative performance of the AI-based auto-captioning for native and non-native speakers. Surprisingly, many technical words that we expected to be missed were accurately captioned (Table 1). Some words and phrases were missed in some, but not all, of the repeated tests. For example, while the phrase *'hinge zone'* comprises common English words, the captions sometimes made this unfamiliar phrase into a single word. Repeating each test at least three times allowed us to assess the variability of performance due to internet quality and other fluctuations. Only six words from the two passages were never correctly captioned with the AI-based auto-caption using the American English recorded under typical sound conditions (Table 1). Words that were missed much of the time for American accented English were missed more often with non-American accented English recordings.

When flummoxed, *Google Slides* captioning, at the time of our testing, would sometimes omit large chunks of the text whereas *Microsoft PowerPoint* mis-guessed a few words. This difference accounts for the larger range of WER for *Slides* captions in Fig. 2. Otherwise, the performance of *Microsoft PowerPoint* and *Google Slides* AI-based captioning was similar under most of the scenarios tested. While analyzing recordings of different accents, we noticed that some words, such as Variscan, were learned by the AI-based captioning and later recognized by the English

recording, yielding 2% improvement in WER. Our experience suggests that jargon can be learned if the AI-based software hears the word in different ways. These codes are updated all the time and might in the future also yield improved caption performance with consistent recognition of jargon placed within the slides/notes.

We tested the impact of audio quality by added background noise and reducing the sound level of the American accented English. The tests show that poor sound quality has a dramatic impact on the quality of the captions (Fig. 2). The WER with poor sound quality reached levels of *YouTube auto-craptions*, exceeding 20% in some cases.

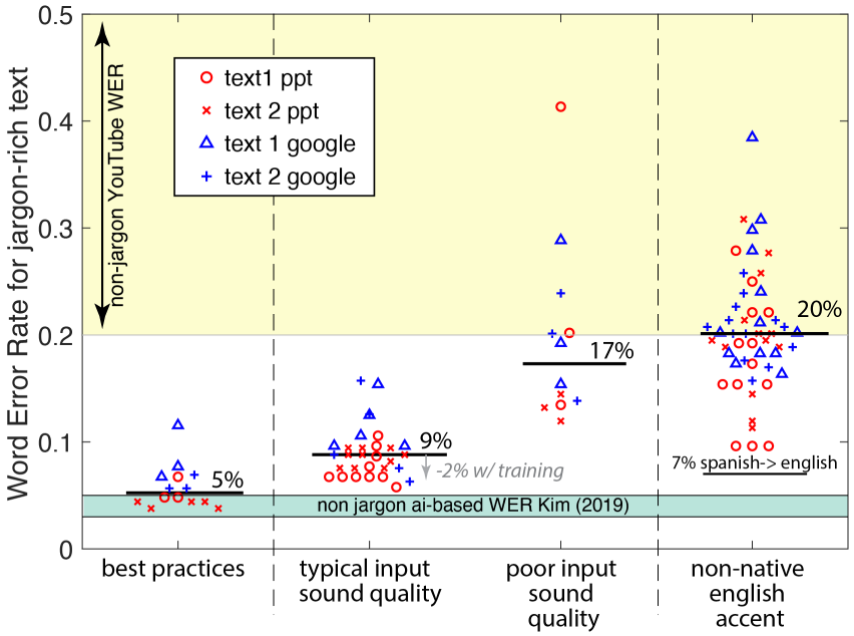

Fig. 2: WER of auto-caption for different settings. The best performance is with lapel microphone. Under these conditions the WER approaches that of non-jargoned text. Poor sound quality and non-native English accents decrease the quality of the AI-based auto-captions for both Microsoft PowerPoint and Google Slides. The % report median for each set of tests.

The WER from recordings of several different people with non-native English accents shows that accents strongly decrease the quality of captioning. *Microsoft PowerPoint* allows the user to choose among several variants on English accents, such as UK and Australian, that were not tested in this investigation. Presumably, if one spoke with an Australian accent with this accent setting chosen, the performance would be similar to that presented here of American accented English (Fig. 2). *PowerPoint* also provides captioning of an extensive set of languages. In a limited test we found that spoken Spanish to Spanish captions performed as well as spoken American English to English. *PowerPoint* also provides translation from one spoken language to another captioned language. We found that the WER for captioning of spoken Spanish to captioned English (~7%) was less than most of the non-native English recordings tests here and the resulting captions missed many of the same jargon presented in Table 1. Some non-native English speakers may find reasonable WER if they use *PowerPoint* translation feature and speak in their native language, allowing the software to translate the captions into another language.

Best Practices

Implementing AI-based auto-caption to live presentations using *Microsoft Powerpoint* or *Google Slides* is straight-forward and yields acceptable quality captioning. Our findings highlight the following best practices.

- Implement AI-based auto-captioning directly within the presentation software. Then folks don't have to run a separate transcription service and switch attention between the presentation to the transcription.
- Speak slowly and clearly. The tests in Fig. 2 for American accented English were from recordings spoken at a conversational pace (average WER of 7.5%). When the same speaker spoke more intentionally, the WER dropped to <6%. The geologic jargon was still missed, but the captioning caught nearly all of the non-jargon words when the speaker pace was slowed.
- Practice with the presentation software prior and see which words are typically missed with your accent. Adding that missed jargon within the text of the slide ensures that the audience can see what the word should be and understand your message. As you repeat jargon in different ways, the AI-based captioning may learn this new word.
- Use an external microphone to improve audio quality. In our tests, using a lapel microphone had the biggest impact on the WER regardless of situation.

Following these best practices of speaking intentionally with a good quality microphone, the WER for the two passages decreased to ~5% over several recordings, a reasonable rate for jargon rich material (Fig. 2). Some jargon that was often missed was captured accurately using these best practices and eliminated other errors from blended and missed words.

We should note that someone with a disability may specifically request a human captionist for live presentations because they provide more accurate captions. Accommodation requests should always be honored. Captionists are expected to have a word error rate of 1% for non-jargon speech (Besner, 2019). While this level of accuracy is needed for some participants, many of us can benefit greatly with captioning of up to 5% error rate such as provided with AI-based live auto-caption. Always include captioning for your live meetings, workshops, webinars, presentations.

Acknowledgments: The authors thank Alina Valop, Xiaotao Yang, David Fernandez-Blanco and Kevin A. Frings for recording their reading of the two passages. The passages tested are available at XXX.

References and Notes:

Besner, L. (2019). When Is a Caption Close Enough? Retrieved September 10, 2020, from https://www.theatlantic.com/health/archive/2019/08/youtube-captions/595831/

Cooke, M., & Caicedo, A. (20120). How to get the most from live auto caption of presentations. Retrieved September 10, 2020, from https://docs.google.com/document/d/1XyAZh6ODq190tx_x1eZ_IsF_L7_-RLF9nLvOGjTR1h4/edit

Kawas, S., Karalis, G., Wen, T., & Ladner, R. E. (2016). Improving real-time captioning experiences for deaf and hard of hearing students. In *Proceedings of the 18th International*

ACM SIGACCESS Conference on Computers and Accessibility (pp. 15–23).

Leduc, J. (2019). The Difference Between YouTube's Automatic Captions, DIY Captions, and 3Play Media Captions. Retrieved September 10, 2020, from https://www.3playmedia.com/2019/02/04/the-difference-between-youtubes-automatic-captions-diy-captions-and-3play-media-captions/

Ofcom. (2006). Television access services review. Retrieved September 10, 2020, from https://www.ofcom.org.uk/consultations-and-statements/category-1/accessservs

Van der Pluijm, B. A., & Marshak, S. (2004). *Earth structure*. *New York* (Second). W W Norton & Company.

Vanderplank, R. (2016). *Captioned media in foreign language learning and teaching: Subtitles for the deaf and hard-of-hearing as tools for language learning*. Springer.

Wang, K., & Liu, H. (2011). Language acquisition with the help of captions. *Studies in Literature and Language*, *3*(3), 41–45.

Weil, A. B. (2006). Kinematics of orocline tightening in the core of an arc: Paleomagnetic analysis of the Ponga Unit, Cantabrian Arc, northern Spain. *Tectonics*, *25*(3).