Captions

Most people have encountered video captioning at some point. Whether on a TV show at a noisy restaurant or on a YouTube video, captions are often around without most viewers giving them much thought. Those who have stopped to consider captions may think of them as an accessibility feature but may not have considered what goes into optimizing their usability. For video creators, however, it is important to understand what captions are, why they are important, and what makes them effective.

What Are Captions?

In the video context, captions can be defined as "one or two lines of text, which represent approximately 1–2 seconds of audio, . . . overlaid on the video screen, which can sometimes obscure video visuals."[1] The captions stay on the screen long enough to be read while moving quickly enough to maintain synchronization with the content of the audio track. It is important to note that this definition refers not simply to spoken content but to all audio content. An important part of captions is translating necessary sound effects and similar audio content, in addition to the spoken language, into text.

As a result, captioning can be a more subjective process than most may realize. This is particularly true in the case of content with noteworthy background sounds where the captioner must decide which background content should be described in the captions and how it should be described. As Sean Zdenek explains it, "Captioning is about meaning, not sound per se. Captions don't describe sounds so much as convey the purpose and meaning of sounds in specific contexts."[2] While for many types of video content, transcribing the contents of the dialogue may be sufficient to capture the full meaning of the audio track, it is important not to fall into the assumption that transcribing dialogue by itself is necessarily sufficient (figure 2.1). To be effective, captions must recreate the experience of listening to the audio content for those who cannot or do not wish to do so. If the captions do not fully represent that content, they will not offer an equivalent experience that is inclusive for those who need or prefer to use captions when viewing video content.

Captions and Subtitles: What's the Difference?

In the United States, the term *captions* typically refers to text that represents the audio in the same language as that audio content, while *subtitles*, on the other hand, refers to text that translates the dialogue into another language. Unlike captions, subtitles typically do not include a textual representation of sound effects and other nonspoken audio content because there is an assumption that the primary users for subtitles will be hearing users who do not fully understand the language but are otherwise able to perceive the audio elements of the video.

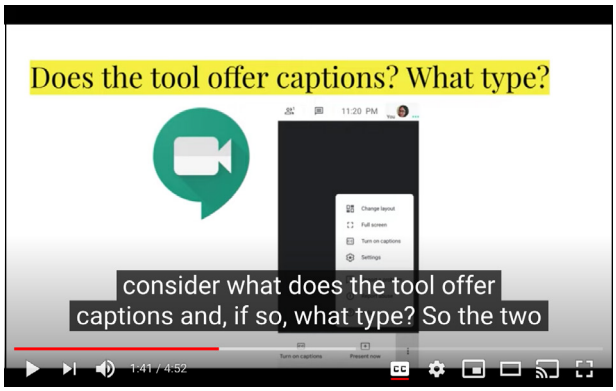

Figure 2.1
Example of closed captions

This distinction is not standard around the world. In many languages and countries, including in many cases in the United Kingdom, the term *subtitles* is used to refer to both types of textual representation of audio content. In these cases, the distinction between the terms is sometimes indicated by referring to text that translates the audio from one language to another as *interlingual subtitles*, while referring to content in the same language as *intralingual subtitles*. In some formal circumstances, this terminology can also be used in the United States, though it is significantly less common. While interlingual subtitles can fulfill some of the purposes of intralingual subtitles or captions, they are not frequently created with accessibility in mind and therefore likely will not fulfill the primary purpose of captions.

Open Captions versus Closed Captions

Captions can be displayed in one of two manners: (1) *Open Captions*, which are permanently visible on the video and cannot be removed, or (2) *Closed Captions*, which can be toggled on or off at the option of the user. Open captions are sometimes also known as *hard-coded captions*, *baked-in captions*, or *burnt-in captions* because they are integrally a part of the video content. One common use of open captions is in live performances, such as theatrical performances, where the audio content is captioned on a screen often above the stage. One advantage of open captions is that the video will always be accessible to those who are D/deaf or hard of hearing without the need for the person controlling the video player to turn on captions. This can be particularly useful in environments such as classrooms or conferences, where the person playing the video may not be aware of the needs of all audience members. In an online environment, open captions also do not require that the video player be compatible with captions. At this point, there is increasingly widespread support for captions in online video players. However, some platforms, such as Instagram at the time of this writing, do not have support for captions in their video players. This means that captions that users can opt to turn on or off are not possible. Instead, the only way to offer captions is to embed open captions in video content before uploading it to the platform, which can be done with many different video creation and editing tools. A disadvantage of open captions is that they can be distracting for users with certain types of disabilities and in certain settings.

Closed captions are the version of captions that most people probably think of when the term is used because they are prevalent both online and off. They are often denoted by one of two symbols, either two *C*s next to each other (figure 2.2), sometimes surrounded by the outline of a screen or television, or the "slashed ear" symbol, which is an icon of an ear with a line through it. This second symbol is also used to indicate services for those who are D/deaf or hard of hearing more generally, but in some parts of the world, or some contexts, it can specifically indicate closed captions. The primary advantage of closed captions is that they allow the viewer to decide whether or not captions are displayed for each individual video based on their specific needs. Many video players also allow users to set persistent preferences if they regularly use captions. The disadvantage of closed captions is that users may not realize they are available, may not know how to turn them on, or, in the case of group viewing of a video, may not realize that some viewers require or prefer captions. Though less of a problem with modern online videos and on televisions in the United States, another potential disadvantage of closed captions is that they require a compatible player to display the captions.

Figure 2.2
Closed caption symbol

A Brief History of Video Captioning

When films initially emerged, they were silent and were inherently accessible to those who could not hear. However, with the emergence of sound films, an access problem arose that was not meaningfully addressed for some time. In 1958, a law was passed to establish a Captioned Films for the Deaf program that loaned captioned films to groups of D/deaf and hard of hearing viewers.[3] Eventually, captions moved to the small screen. The first instance of open captioned content on television was rebroadcast episodes of Julia Child's *The French Chef*, which started on WGBH in 1972, followed by the debut of open captioned rebroadcasts of *ABC World News Tonight* on the same channel.[4] It was not until 1980 that closed captions debuted on American television, and by the late 1990s over 500 hours of captioned programming was broadcast each week.[5] Though it may be surprising that open captions preceded closed captions by so many years, this is because it was not until the Television Decoder Circuitry Act of 1990 that all televisions with a thirteen-inch or larger screen in the US were required to have the technology necessary to support closed captions.[6] Prior to that, those who wanted to access closed captions needed external equipment, which limited the reach of the technology, particularly in public spaces where captions are often seen today, such as airports and restaurants.

With the advent of online streaming video, captions became important for a new type of video that offered access to new types of content. One of many ways in which online video is different from television programs is that more of it is created by individuals or institutions that may not have much experience or expertise in video creation, which likely contributed to the lag in captioned video online. However, there were other significant factors, including a lack of support for closed captions on online video platforms such as YouTube, which did not have support for closed captions until 2008.[7] In recent years, lawsuits have helped to expand the availability of captions, including influential lawsuits filed by the National Association of the Deaf against Netflix, Harvard, and MIT, which helped advance online captioning significantly.[8] Though the availability of captioned online videos has increased, there are still many uncaptioned, inaccessible videos available online and even entire platforms that either do not support captions or offer only minimal access for users.

Why Is Captioning Important?

From an accessibility point of view, captioning is vital for users who are D/deaf or hard of hearing, and these users should always be the top priority when designing captions and media player support for captions. Given that about 15 percent of the population of the United States has at least some difficulty hearing, this constitutes a significant audience. Including captions makes video content more inclusive for these users. It also fulfills basic legal requirements that many organizations must meet, especially government and educational institutions.

However, D/deaf and hard-of-hearing users are hardly the only audience for captions. Many different users find captions useful in many different settings, including

- users who process information better through text;
- users who need or want to watch videos without the audio on, whether due to their setting, such as when watching in a library, or to not disturb others around them;
- users for whom the language of the video is not their primary language, particularly when subtitles are not available;
- users watching videos with speakers who mumble, with unclear audio tracks, or with speakers with multiple accents;
- users learning new terms or concepts that might be easier to comprehend either through text or a mix of text and audio; and
- users who are learning to read.

Data shows that captions are popular in these and other situations. A 2019 study of consumers 18 to 54 years of age by Verizon Media and Publicis Media found that 80 percent of those using captions are not D/deaf or hard of hearing but are actually using captions for another reason.[9] The same study also found that 80 percent of respondents said that the presence of captions made them more likely to watch a video.[10] Other studies have also found that captions impact viewership. A study by 3Play Media and Discovery Digital Networks (DDN) found that there was an "overall increase of 7.32% in views for captioned videos" on DDN's YouTube channel.[11] A nationwide study of students at institutions of higher education found that 70.8 percent of surveyed students who did not have any type of hearing difficulty used closed captions when watching at least some of the videos associated with their courses.[12] No matter the setting, it is clear that many users prefer to use captions.

Beyond their popularity, captions also offer benefits for virtually all users. In fact, a 2015 review of the literature found that over 100 empirical studies had shown benefits of captions for users of many ages and in many scenarios.[13] In educational settings, captions have been shown to be particularly useful. A study of caption use in language learning classes found that captions "result in greater depth of processing by focusing attention, reinforce the acquisition of vocabulary through multiple modalities, and allow learners to determine meaning through the unpacking of language chunks."[14] Beyond language learning, captions have been demonstrated to have notable benefits for students at many different levels, from elementary school to college.[15] While it is vital that users with disabilities remain the primary focus when designing video captions, it is equally clear that captions will be beneficial for many other users.

How Are Captions Created?

There are three primary ways that captions are created. Until recently, captions were almost always created by an individual typing up captions for the content during or after creation of the film or video. These individuals are sometimes referred to as stenocaptioners if they use stenography equipment for the process. This method can be used for both prerecorded content and live content. However, another way that captions can be created in some platforms is by typing up or uploading an existing script of the dialogue, either with time stamps built in or using a tool that is capable of detecting sounds and automatically lining up the captions. Using more recent technologies, captions can also be created using artificial intelligence (AI) word recognition. Well-known applications such as PowerPoint, Google Docs, Zoom,

and YouTube have automatic captioning features built in to their programs. Though the idea of automatically generated captions is appealing, the accuracy of these automated tools still lags behind the accuracy that can be achieved by human-created captions, particularly when the audio is unclear for any number of reasons, from recording standards to the level of enunciation of speakers. Recent research demonstrates that this issue persists in particular in videos with technical terminology.[16] For many institutions, this automated approach to captioning must be combined with a human review after the fact to find and fix any errors. However, automatically generated captions are increasingly integrated into video conferencing tools to support captioning live events. Skilled stenocaptioners can provide more accurate real-time captions in many cases, and many of these platforms also provide an option for integrating captions created in this manner.

Caption Accuracy

Though it may seem obvious that captions should be an accurate representation of the audio content, views on the best approach to accuracy have changed over time. Initially when captions were aired on television, they intentionally did not exactly represent what was said in the video and instead edited the content to ensure that the captions were written at a lower reading level, a fact that some researchers have argued was accepted at that time at least in part because "deaf people were so delighted to have captions that they accepted almost anything thrown on the screen."[17] Over time, this model shifted significantly so that it is now much more common for captions to be defined as the "verbatim translation of spoken dialogue."[18] In fact, best practices are generally to offer 99 percent accuracy, a level that is offered by many vendors that provide commercial transcription and captioning services. This high level of accuracy is needed to ensure that the video is comprehensible for users who have no access to the audio track. For this reason, workflows that involve automated captions generally also need to incorporate a review to ensure the accuracy of the generated captions.

Though accuracy is vital, the meaning of accuracy can be more complicated than it might seem at first. One often-overlooked fact about captions is that they are, to at least some degree, subjective. While they should strive to recreate the sound of the video content, the final product may well differ, most particularly when there are non-dialogue elements integrated into the audio. In fact, there will often be more than one official set of professionally produced captions for a single movie or TV show that is released in different settings, such as a television broadcast and a DVD release. This is because captions are intended to translate the full spectrum of the sounds that are part of the video. They are meant to convey not only the meaning of dialogue that is unclear, and therefore subject to interpretation by the captioner, but also the important background sounds and sound effects, and in some cases a descriptor of a character's emotion. Any sound that conveys meaning is integrated into the captions for a video. As Zdenek argues, in at least some contexts, "captioners not only select which sounds are significant, and hence which sounds are worthy of being captioned, but also rhetorically invent words for sounds."[19] It is also important to note that, though the modern best practice is generally to caption all spoken words, captioners in some cases may be required to also rephrase or condense spoken content to reasonably be read by viewers during the duration of the relevant video content. All of these factors mean that some experts recommend employing experts to create captions for videos used in educational settings, though of course this has associated costs.

Best Practices for Caption Creation

For those who are interested in creating captions, there are some best practices that can help to ensure that the completed captions offer meaningful access for users:

- Accuracy is vital. Strive for 99 percent accuracy for prerecorded captions. When providing captions for a live event, strive for maximum accuracy and, if a recording will later be provided, correct the captions before providing access to the recorded video. When using automatic captioning features, check and correct captions as necessary to achieve 99 percent accuracy.
- Avoid obscuring important content in the video with the captions.
- Ensure that the font size of the captions is large enough to be comfortably read even by those with low vision. Generally, the font size recommended for accessibility is no smaller than 16 points, and captions should be one or two font sizes larger than that. However, that will vary depending on the size of the video, and not all platforms will allow the caption creator to select the size of the font.
- Choose a font that is very readable. Generally, sans serif fonts such as Arial, Helvetica, or Verdana are preferred for this purpose, though not all platforms offer multiple font options.
- Select a font color that will be high contrast compared to the video content if the captions will be overlaid over the video or high contrast compared to the background if the captions will be on a solid

background immediately below the video content. If the platform being used offers only a single caption color, it is important to consider where the captions will appear on the screen and attempt to ensure that the background behind the captions will offer a high contrast backdrop for the text.

- If at all possible, allow users the flexibility to select between several fonts, font sizes, and font colors to find the settings that work best for them. This feature is not supported by all video players, but it should be offered when supported.
- Censor only content that is censored in the audio track. For example, if profanity is bleeped out in the audio track, it should be similarly censored in the captions, but if it is not bleeped out in the audio track, it should not be censored. Content that is censored according to this model should be reflected either by replacing some letters in the middle of the word or by simply typing *[expletive]* in place of the word.
- Limit the number of words and characters on the screen at any time to ensure that the text is readable.
- Caption synchronization is important. The text on the screen should be closely synchronized with the audio track. In the case of prerecorded video content, this synchronization should be exact. When creating captions during a live event, complete synchronization is not possible, but synchronization should be as close as possible.
- The text should remain on the screen long enough to be readable. In the case of fast-moving dialogue, this may at times require some abridging and editing of the content. However, this should be done only when absolutely necessary as verbatim captions are preferable.
- Sounds indicating pauses or serving as fillers, such as *um*, *ah*, *hmm*, or similar, can be omitted as long as their omission does not prevent those reading the captions from understanding the meaning of the content. Similarly, if a speaker misspeaks or repeats a word, this may also be omitted if it does not impact the meaning of the content.
- Sound effects should be captioned in addition to dialogue. Similarly, captions should indicate when music is playing and should caption lyrics, particularly if they are relevant to the meaning of the content.
- In the case of dialogue where it is important to know who is speaking and this may be unclear to those viewing the video without sound, the captions should indicate this information. For example, if dialogue is spoken by someone off screen, this should be indicated.

Captions are a vital element of accessibility. While they are increasingly found in videos both online and offline, unfortunately, many videos still lack captions. The advent of automated captions on platforms such as YouTube has increased their prevalence, but issues of accuracy remain. In order to provide an equitable and usable viewing experience regardless of access to the audio track, it is important to incorporate accurate captions into all video content.

Notes

1. Raja S. Kushalnagar, Walter S. Lasecki, and Jeffrey P. Bigham, "Captions versus Transcripts for Online Video Content," in *W4A '13: Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, 1, http://www.cs.cmu.edu/~jbigham/pubs/pdfs/2014/captionvstranscripts.pdf.
2. Sean Zdenek, *Reading Sounds: Closed-Captioned Media and Popular Culture* (Chicago: University of Chicago Press, 2015), 8.
3. Captioned Films Act of 1958, Pub. L. No. 85-905, 72 Stat. 1742 (1958).
4. Carl Jensema, Ralph McCann, and Scott Ramsey, "Closed-Captioned Television Presentation Speed and Vocabulary," *American Annals of the Deaf* 141, no. 4 (October 1996): 284.
5. Carl Jensema, "Viewer Reaction to Different Television Captioning Speeds," *American Annals of the Deaf* 143, no. 4 (October 1998): 318.
6. Television Decoder Circuitry Act of 1990, 47 U.S.C. §§ 303(u) and 330(b) (1990).
7. YouTube, "YouTube Captions and Subtitles," posted September 22, 2008, YouTube video, 1:36, https://www.youtube.com/watch?v=QRS8MkLhQmM.
8. National Association of the Deaf, "NAD Files Disability Civil Rights Lawsuit against Netflix," June 16, 2011, https://www.nad.org/2011/06/16/nad-files-disability-civil-rights-lawsuit-against-netflix; National Association of the Deaf, "Landmark Agreements Establish New Model for Online Accessibility in Higher Education and Business" (news release), February 18, 2020, https://www.nad.org/2020/02/18/landmark-agreements-establish-new-model-for-online-accessibility-in-higher-education-and-business.
9. Verizon Media, "Make Noise with the Right Digital Video Captioning" (infographic), April 2019, https://b2b.verizonmedia.com/c/verizon-media-sound--1?x=vOJKbY.
10. Verizon Media, "Make Noise."
11. 3Play Media, "Discovery Digital Networks," accessed October 24, 2020, https://www.3playmedia.com/why-3play/case-studies/discovery-digital-networks.
12. Katie Linder, *Student Uses and Perceptions of Closed Captions and Transcripts: Results from a National Study* (Corvallis: Oregon State University Ecampus Research Unit, October 2016).
13. Morton Ann Gernsbacher, "Video Captions Benefit Everyone," *Policy Insights from the Behavioral and Brain Sciences* 2, no. 1 (2015): 195–202.
14. Paula Winke, Susan Gass, and Tetyana Syodorenko, "The Effects of Captioning Videos Used for Foreign Language Listening Activities," *Language Learning and Technology* 14, no. 1 (2010): 81.
15. Faye Parkhill, Jiliane Johnson, and Jane Bates,

"Capturing Literacy Learners: Evaluating a Reading Programme Using Popular Novels and Films with Subtitles," *Digital Culture and Education* 3, no. 2 (2011): 140–56; Aaron Steinfeld, "The Benefit of Real-Time Captioning in a Mainstream Classroom as Measured by Working Memory," *Volta Review* 100, no. 1 (1998): 29–44.
16. Tharindu R. Liyanagunawardena, "Automatic Transcription Software: Good Enough for Accessibility? A Case Study from Built Environment Education," in *European Distance and E-Learning Network (EDEN) Proceedings: EDEN 2019 Annual Conference, Bruges, Belgium*, ed. Airina Volungeviciene and András Szűcs (Budapest, Hungary: European Distance and E-Learning Network, 2019), 388–96.
17. Jensema, McCann, and Ramsey, "Closed-Captioned Presentation Speed and Vocabulary," 285.
18. John-Patrick Udo and Deborah I. Fels, "The Rogue Poster-Children of Universal Design: Closed Captioning and Audio Description," *Journal of Engineering Design* 21, no. 2–3 (2010): 207.
19. Zdenek, *Reading Sounds*, 1.