A Tutorial for Video in Spoken Language Documentation

Abstract

Spoken language always goes along with meaningful visible behavior, such as gesture and eye gaze. But while language use is multimodal, published recommendations and formal training in spoken language documentation tend to focus almost exclusively on the audio part of the signal. Therefore, this tutorial provides a practical guide to using video as part of a spoken-language documentation project. We motivate why these projects should consider recording video, and we then describe the equipment needs, recording setups, and post-processing workflow required for collecting transcribable video. We also discuss the unique ethical/privacy concerns raised by video recording and archiving. Overall, our goal is to centralize and formalize the recommendations about video that have long circulated in oral form, or as grey literature, in documentation circles.[1]

1 Introduction

Seyfeddinipur & Rau (2020: 517) close their recommendations for video use in language documentation with "the future is bright and exciting and confusing … so watch this space." Since the goal of language documentation is multipurpose records of language in context (Himmelmann 1998), it is important to create linguistic collections which are robust, which can be repurposed for different audiences and research questions, and which can be archived securely.[2] In this tutorial, we cover the key decision points for equipment choice, recording, and processing for multimodal linguistic video documentation: why one should do video documentation, how to do it, and why to do it that way. Following preliminaries on the benefits of recording video (in §2), we summarize the decision points that need to be made as part of recording with video. We then give step-by-step instructions for choosing equipment, setting up shots, postprocessing video data following recording, and archiving.

[1] [Acknowledgements redacted for anonymous review]

[2] Both Seyfeddinipur & Rau (2020) and Berez-Kroeker et al. (2023) briefly discuss Augmented Reality recordings as a potential for documentation. Pentangelo (2020) puts this into practice. We focus here, however, on traditional camera-based techniques. This tutorial is about video documentation in person, not for fieldwork conducted remotely.

Berez-Kroeker et al. (2023: 200) note that “guidance on audio standards and equipment selection for language documentation is easy to obtain (e.g., Bowern 2015)”. The same cannot be said for video standards. Echoing Seyfeddinipur & Rau (2020), Berez-Kroeker et al. point out that available advice on video is usually geared towards documentary film makers or YouTube content creators, and video for language documentation has different requirements. Yet video documentation is seldom taught in field methods classes, and where it is taught (e.g., for signed language documentation), it is not taught by researchers who focus on sound.[3] Perhaps as a result, the results from video documentation can be less than satisfactory.

Besides, skill in using audio will not transfer directly to video recording. For instance, audio devices need to be close to the speaker in order to pick up their voice. Video devices need to be relatively far in order to keep the participants in the camera frame even if they stand up or make large pointing gestures. Reconciling the competing needs of the audio and video streams can be technically challenging.

Researchers using video for aural-modality language documentation should therefore learn from signed language documentation. Hochgesang & Fenlon (2022) is a very useful overview of documentary linguistics for corpus creation (centered on signed languages and deaf communities); in particular, Hanke & Fenlon’s (2022: 35ff) corpus creation chapter provides advice about equipment and camera setup. They advocate using a studio setup, particularly because of the need for consistent backdrops to make sign data more visible. Fenlon et al. (2015) describe their setup for gathering data for the BSL corpus; they also discuss how the recording setup affects data gathering, such as the impacts of a studio setup on naturalistic data collection. The most comprehensive descriptions of signed language documentation and video recording come from projects on signed languages in urban, Western environments where a studio setup is feasible.[4] While a studio setup for recording audio and video for spoken languages might be advantageous in some circumstances, it is not usually possible or advisable, since one of the reasons for making visual documentary recordings is to provide further information about language use in its context.

We should also learn from anthropology, where visual documentation is far more common and has a much longer history. As Everri et al. (2020: 68) put it, “visual methods have been essential in ethnography from the start.” Consider Mead & Bateson’s (1951) film “Bathing Babies in Three Cultures”, for example. Ethnographic film is a subdiscipline within sociocultural anthropology and there is a clear literature on both the technical aspects of video and the theory:

[3] As Tsikewa (2021) points out, field methods classes are the first exposure that students receive to language documentation, and what we teach in such classes is a clear signal of what is prioritized in the field. Likewise, field methods textbooks seldom provide extensive instruction on video. Bowern (2015), for example, discusses “audio and video” recordings but most of the explicit information is geared towards the audio component.

[4] Goico (2019: 17, 87) is one example of a naturalistic setup (albeit in classrooms) for recording homesigners in Peru.

how to formulate research questions that can then be studied and documented visually, and how different choices made by researchers create "focused" elements in contrast to "the broad panorama of a culture" (Woermann 2018: 459).[5] At the same time, visual anthropology has different goals than language documentation, making discipline-specific advice necessary.

2 Why record video when focusing on sound

2.1 Advice against using video

Researchers often resist using video in their spoken language documentation projects. However, we believe that reasons in favor of recording video far outweigh the reasons against. Recording video has at least three advantages, which we discuss in more detail below in §2.2. First, visual information links the speech signal to the surrounding context, which is often crucial for understanding what is going on. Second, visual information (e.g., from co-speech gestures) elucidates grammar by making transcription easier and captures information that is relevant for grammatical analysis. And third, when done correctly, the visual signal extends the usability of general-purpose language documentation, which is a strong reason for doing language documentation in the first place.

Furthermore, the arguments against using video are not particularly compelling or unique to the visual channel. For example, some advice suggests that the presence of a video camera increases the artificiality of the situation; that is, it increases participants' self-monitoring and alters behavior. That is true, but it is true for audio recording too. Research participants usually become more comfortable with video fairly quickly, just as with audio. "Icebreaker" exercises can be an important component of beginning a structured (or semi-structured) field recording session for this reason. Some research participants may be uncomfortable with video, but others probably will be happy to be recorded, just as participants vary in their degree of comfort with technology more generally. And as Seyfeddinipur & Rau (2020: 503-504) point out, norms around appropriate individuals and circumstances to record will vary across cultures. That is, video documentation is a point to discuss as part of consent to work with individuals and a community, along with all the other points around collection, dissemination, and archiving of materials that linguists discuss across the life of a research project. For video as for these other points, linguists should be aware that their own views about what they are (un)comfortable with might differ from those they are working with.

Another line of argument is based on recording quality. Poor-quality documentation impedes analysis; therefore, given that fieldwork already involves acquiring a vast array of skills (cf. Bowern 2015's "many hats"), perhaps it is preferable to focus on better audio recordings than

[5] See also Mitsuhara & Hauck (2022), Dimmendal (2010); and Moriarty (2020). Enfield (2013) provides an overview of considerations for documentation of linguistic research involving the body, including ethical considerations for video.

more multimodal information, but with worse quality.[6] We agree that poor-quality documentation impedes analysis, but the solution is not to restrict the modality of recording. Hanks' (2009) arguments for incorporating video in analysis of deixis is a great example. Even if fieldnotes or audio recordings are extremely faithful, they only preserve information that the researcher noticed or deemed important at the time, which means they can obscure crucial parts of the spatial or interactional context for analysis of deictics. Visual recordings, on the other hand, are less intrinsically selective in attention and allow for a better understanding of the grammatical category.

Researchers also often cite budget as an obstacle to recording video. While the semi-professional video cameras recommended by archives (e.g., PARADISEC[7] and ELAR[8]) cost well over $1000 USD, many mid-range or 'prosumer' cameras are closer to $600-$800 USD (or can be bought for even less refurbished). Peripherals and additional storage for video can be purchased for as little as $200-$300 more. Even for a dissertation grant or a small (<$10,000) grant, this is a reasonable addition, especially compared to the cost of travel and the cost of other equipment. Archiving video does also incur extra storage costs to the archive, but how large these costs are and whether they are passed on to the researcher varies enormously across archives. We provide recommendations for video setups in a range of budgets in §3.4, including budgets under $500 USD.

People worry about the logistics of backing up and post-processing video data in field situations too. This was justified in the 2000s and early 2010s when (for example) some cameras still recorded to videotape, forcing researchers to return from the field to have tapes digitized before annotation (e.g., Dingemanse 2011: 9ff). However, because of declines in the cost of storage space and improvements to video technology, recording and processing video is now much simpler. Additionally, camera technology and video standards have stabilized over the past ~10 years, meaning that researchers no longer need to make significant changes to their video workflow or equipment year-on-year. For example, the first author has used the same video workflow described in §4 since mid-2018.

Last, transportation and electricity needs are also sometimes raised as obstacles to recording video. It is absolutely true that, even with improving video technology, semi-professional video cameras are still bulky and heavy. For example, a kit consisting of a semi-professional camera, external cardioid microphone, other peripherals, and hard-sided camera case can weigh up to 10kg and require its own airline carry-on. However, as with the issue of price, there are many options besides semi-professional cameras which are lighter and easier to transport in a remote

[6] Compare Henke & Berez-Kroeker's (2016: 446–447) summary of Nathan (2009).

[7] https://www.paradisec.org.au/resources/downloads/

[8] https://www.elararchive.org/dk0000

field setting. We provide recommendations for a variety of sizes and weights in §3.4. The same section also includes advice for running cameras on limited electricity.

Of course, field sites, situations, and communities vary extensively across the world, and with it, the ease and advisability of making extensive audio and video recordings. Our point here, however, is that the reasons for not recording should be the actual constraints brought on by fieldwork, rather than issues that have straightforward logistical solutions.

2.2 Benefits of recording video

Video helps document the situation. That is, it provides a record of the surroundings in which the linguistic utterances are made, and that information helps us understand more generally what is going on. As Deal (2015: 157) notes, linguistics relies on reasoning from a linguistic behavior (an utterance, a response to a stimulus in an experiment, and so on) to meaning: understanding why people say what they do when they do, and what information they do (or do not) convey by the choices they make. Labov (1972) describes language variation as studying the choices that individuals make—given their linguistic repertoires and the flexibility of their systems, they always have choices—and those choices can be situational. There are many features of the context which we only know from visuals and so they should be documented along with the auditory stream.[9]

Video vastly improves the ease of the transcription of data, particularly for interactional data where one must make sense of who is talking, who the utterance is addressed to (individual participants, collectively, for example), and how the individuals respond. A detailed example from Aboriginal Australia is given in Blythe et al. (2018), which demonstrates that speaker selection shows a range of tacit and explicit embodied signals of turn-taking. Video is not only a record of who is present. It can also document the other activities occurring in the situation. For example, imagine someone telling a narrative, where they break off in the middle to make a parenthetical comment to someone else in the room. The reason for the comment may not be recoverable from the audio alone when the recording is analysed later on. Some of this information might be noted in situational metadata, but Hanks' (2009) point is also relevant here: such material is intrinsically highly selective when its recording is contingent on the researcher. For example, it might note that another person was present, but not whether (or how) they were engaged with the session.

Resolution of pronominal reference or reference to actions is much easier with video. In (1), for example, three speakers of the Amazonian isolate Ticuna—KGW, AYM and EWI—are discussing how AYM should treat a sore inside of her mouth (this example is reproduced from

[9] Nathan's (2009; 2010) notion of audio as "evidence" for the derivative textual rather than as performance and primary source of analysis also applies to video documentation.

Skilton 2019: 188-189). KGW first suggests treating it with clotrimazole, an antifungal cream. She calls to her daughter to bring them some (line 1) and then to bring a mirror (line 3). However, she quickly changes plans and instead tells AYM to apply hot ashes to the sore (line 4), pointing at the hearth in the back of the room. It is only because of KGW's pointing gesture, visible on the accompanying video recording, that we can tell that the demonstrative in line 4 refers to the ashes, rather than the clotrimazole. In line 5, KGW then demonstrates with a gesture how AYM should apply the ashes. Again, it is only the video that makes clear that line 5 involves reference to an action, rather than an object.

(1) California Language Archive item 2018-19.024, file tca_20170527_disc_video_007_archive.mp4, 8:13[10]

1. KGW: Andrea! e^{3}ĩ^{1}ka^{5}ʔ, clotrimasol nu^{2}a^{2} na^{1}ʎe^{43}

Andrea	e^{3}ĩ^{1}ka^{5}ʔ	clotrimasol	nu^{2}a^{2}	na^{1}=ʎe^{43}
personal.name	let's.see	SP:clotrimazole	here	IMP=transport

'Andrea! Come on, bring the clotrimazole here.'

2. EWI: tu espejo

(Spanish) 'Your mirror?'

3. KGW: m^{31}, ɲu^{1}ʔũ^{4}tʃi^{5} ɟi^{31}ʔe^{2}ma^{4} espejito nu^{2}a^{2} na^{1}ʎe^{43}

m^{31}	ɲu^{1}ʔũ^{4}tʃi^{5}	ɟi^{31}ʔe^{2}ma^{4}	espejito	nu^{2}a^{2}	na^{1}=ʎe^{43}
INTJ	also	DEM(I)	SP:mirror(I)	here	IMP=transport

'Mm, bring that mirror here too.'

((AYM and KGW gazing at each other))

4. KGW: ŋe^{3}a^{2} ma^{4}ã^{2}i^{5}ra^{1}

ŋe^{3}a^{2}=ma^{4}ã^{2}=i^{5}ra^{1}

DEM(IV)=COM=first

'First (do it) with that stuff (=ashes)'

((KGW pointing with right hand at fire))

((KGW gazing at pointing target, then at AYM))

5. KGW: ɲa^{4}a^{2} dɨ^{1}ʔ

ɲa^{4}a^{2}	dɨʔ
DEM(IV)	look!

'(Like) this, look.'

((KGW makes an iconic gesture: dabbing inside of mouth with index finger of right hand))

[10] This example uses IPA for segments, raised numerals for tones, and Leipzig abbreviations. SP = Spanish word. Roman numerals such as 'I' and 'IV' represent noun classes.

((KGW gazing at AYM))

Several different types of language documentation are particularly amenable to video documentation. This includes language use involving material culture or the built environment, or anything involving a practice: that is, not only documentation about the things that people say in a language, but also about those things themselves. Imagine a recorded discussion of weaving, for example, where the speaker is giving instructions but we cannot see which threads are referred to. "Texts" and interactions about the natural environment also benefit from knowing the situation of utterance, for interpreting ambient and locational information (Engman & Hermes 2021).

Attention to the environment applies to structured elicitation along with semi- or unstructured activities. For example, language production in elicitation sessions may change depending on who else is in the room, as language users accommodate to speech styles, which adds another dimension to the task but the researcher may be oblivious to (particularly if they are not very familiar with language use outside a structured elicitation context). Any language use that involves multimodal channels (co-speech face and hand gestures, for example, as discussed below), whether elicited or conversational, needs visual documentation for clear interpretability.

Video is important for language reclamation, for many of the same reasons that it is important for linguistic analysis. Several authors (Zabulis et al. 2024 on multimedia dictionaries; Degai et al. 2023 on Itelmen, for example) stress video along with audio for understanding processes that have a visual component; understanding the full context of the utterance; and in particular, for making connections between language and other aspects of social life. That is, visual information is a crucial link that connects information contained in the speech signal to its meaning, not just ancillary evidence that supports a transcription (Nathan 2010). This is particularly important for language learners and connectionist language reclamation which seeks to support language as one of many important threads of culture (Leonard 2017), or which defines language as including multimodal information (Hermes et al. 2023). Besides, as many of the papers in Turin et al. (2011) discuss, oral literature has a performance component that is best understood with visual information.

Spoken language is usually interpretable without a visual channel, but co-speech gestures and head movements contribute key information to the discourse context and shed light on other aspects of grammar. This is easily apparent for work on deixis and demonstratives (Hanks 2009) but is also true for other areas of grammar, such as aspect (Ko & Laparle 2022) and serial verb constructions (Defina 2016), for example.[11] Some audio information is easier to transcribe with additional visual cues to articulation. Lamino-dental and apico-dental consonants, for example, are similar acoustically, but are more easily distinguishable visually, since the lamino-dental set

[11] See Church et al. (2017) for a broad overview.

often has a visible protrusion of the articulator (that is, the tongue blade is easier to see). Video also captures relevant information for language documentation of speech sounds, such as degrees of lip rounding.

Finally, video documentation increases the potential future uses for generalist documentation collections. Language reclamation repeatedly shows us that linguistic researchers do not (and cannot) anticipate all the uses to which records of language will be put. Therefore in addition to collecting data on the specific questions that interest them, they should do as much as they feasibly can to make sure that the records they make can be used for other purposes. There are straightforward things which increase the use potential of linguistic records (like using decent audio recording even if you are only going to transcribe it and work from the text, saving filler sentence translations as well as the data of primary experimental interest, etc). This is an ethical requirement of anyone who is conducting research in a community where the reason elders agreed to participate was to help pass down the language to future generations.[12]

2.3 Ethical issues in video documentation

Numerous papers discuss ethical considerations in the documentation of languages of all modalities, from several viewpoints (see among others Hochgesang & Palfreyman 2022; Cychosz et al. 2020; Rice 2006; D'Arcy & Bender 2023). Many points related to video documentation will be covered by existing ethical frameworks: the need for informed consent, discussion of where material may be released and for what purpose, and so on. This is worth reiterating given there is no expectation that large AI-mining companies will respect Indigenous intellectual property rights over data hosted on their services (Holton, Leonard & Pulsifer 2022).

The same reasons that we advocate video documentation—the rich recording of the surrounding environment, for example—are also potentially liabilities and increase the risk of recording. They make it much easier to identify people and places (which may be risky, depending on the individuals). Research participants' willingness to have their images archived varies enormously across the world (see e.g., Enfield 2013: 978), and appears to be changing. As the risks of freely available images and video become clearer, people may be less likely to wish to include their images in public corpora. Conversely, as video becomes more widespread with the ever-increasing popularity of sites like YouTube and TikTok, and people become more habituated to placing video online, they may become more comfortable with video documentation for linguistic purposes.

[12] Pentangelo (2020) and Degai et al (2023) are good examples from two different regions. Pentangelo (2020) discusses general purpose documentation for Kanien'kéha but focuses on community priorities and materials for language learners.

Both audio and video documentation may end up recording material that has privacy implications. Linguists have long been warned about accidentally recording and releasing personal information that might not be appropriate for a general audience, and video intrinsically captures more information than audio. Aston and Matthews (2011) discuss one example, where an ethnographic film collection created for personal and teaching reasons ended up documenting aspects of the Sudanese civil war.

3 Step-by-Step instructions

In this section we work through the decisions to make and equipment, software, and skills needed for doing video documentation. We shift to address the reader directly for ease of reading.

3.1 Decision points

Before selecting a camera kit and recording process, you need to answer the following questions.

First, what type of events do you plan to record? How many people will be recorded, how far apart will they be in space, and how much will they be moving? The answers to these questions will determine the technical specifications that you need for your kit. For example, if you plan to record interactions with more than two people, you need to be able to take a much wider shot than if you will only be recording 1-2 people at a time. Similarly, suppose you plan to record events where the participants are moving/physically active during the event -- this often happens in video texts about material culture or the built environment, as well as when recording interaction. In this case, you'll need to mic a bigger region of space and use a less directional mic than if you only recorded videos where the participant was completely stationary. Also, even if you expect to be recording only one event type, fieldwork is filled with unexpected opportunities. You should be prepared to benefit from them, so it's a good idea to equip yourself for a slightly broader range of event types than you actually expect.

If you don't know what to expect, for example because it's your first trip, it's best to equip yourself to be able to record two moving participants, in a domestic-sized space, using a single camera. This setup will still work if you only end up recording one participant at a time, or only record stationary participants. You'll have trouble recording interactions that use certain facing formations (e.g., two people facing each other across a very small space), as well as recording people moving across landscape-scale space. However, you can probably still capture some reasonable quality footage with this setup, and these are less common recording types for language documentation in any case.

Second, will you be recording the audio track on your videos using a microphone connected to your camera ('single source'), or will you be recording with a separate audio device/devices ('multiple source')? As we discuss in much more detail in §3.7, the main advantage of single-source is that you will not need to synchronize the audio and video sources during post-processing. However, if you want to record acceptable quality audio and video with a single source, you will need to buy either a wireless microphone system or a high-quality cardioid microphone plus a mic stand. This increases cost, greatly increases the number of failure points in your setup, and (if you use a mic stand) requires you to transport a large extra piece of luggage. Unless you have access to a reasonably priced and reasonably failsafe wireless microphone system, using multiple sources (§3.8) is much lower risk. Additionally, a multiple-source setup provides backup in case of equipment failure.

Last, will you be recording with a single camera, or with multiple cameras at opposing angles? Multiple cameras allow you to record a larger region of space (see §3.9 on shot composition with multiple cameras), as well as to include multiple angles. This is helpful if you are recording (1) scenes where the participant(s) are extremely mobile; (2) a space that is larger than ~10 square meters; (3) two participants who are directly facing each other (as they then can't both face a single camera); (4) a scene with >2 participants (as they're unlikely to all face a single camera) or (5) a scene where you need two shots of different size (e.g., a close-up shot of a participant's lips or face for visual phonetics and a wider shot of their entire body). Multiple cameras seem to be the standard for in-person sign language fieldwork, even when recording a single primary participant (cf. examples in, among others, Fenlon et al. 2015; Hochgesang 2020; Hou 2016; Schembri 2010). So even if none of the above space considerations apply to you, if you are interested in studying multimodal/visible behavior, it's probably wise to have multiple cameras.

3.2 Selecting a camera

First you must choose a camera. The main considerations in picking a camera should be price, size, and mobility.

3.2.1 Price

You need a camera that fits your budget, but you should not sacrifice image quality for price. Some very-low-budget (under 500 USD) cameras deliver a distorted image. For example, the Zoom Q8 is a popular camera, but has a fisheye lens which emphasises the central focal point of the image area.[13] We are not convinced that it should be considered acceptable quality for analyzing visible behavior. Also, footage with this quality might not be acceptable for distribution in the community, as it is noticeably worse than footage from many smartphones. So if your budget is in this range, it is a better use of money to buy a used camera of higher quality instead—refurbished cameras are sold widely. There is also the option of recording using a device that is not a dedicated video camera, such as a smartphone or tablet mounted on a tripod. This probably will not be more expensive than the Zoom Q8 and will produce significantly better

[13] The Zoom Q8N-4K may deliver slightly better image quality, but is also more expensive.

quality footage. (Such devices could also potentially be left with community members for further documentary work.)

3.2.2 Size

If you are going to be walking significant distances with the camera and kit, or if you would like to avoid using a camera case for security reasons (§3.5) you probably want a camera that you can carry comfortably in a backpack. The same applies if you plan on filming handheld (although you should avoid filming handheld, if possible, for the reasons in §3.5 and §3.9).

It is difficult to fit a semi-professional camera like the Canon XA series into a backpack, and they are also very heavy to use for handheld filming. While these cameras are versatile and have great image quality, that doesn't matter if you are unable to use the camera because you can't carry it with you. If transporting a 10kg camera kit would be a serious problem in your field setting, you should use a smaller prosumer camera, such as the Sony Handycam series. When making weight and luggage calculations, bear in mind that you will be carrying a tripod as well as your camera(s).

3.2.3 Mobility

Conventional video cameras, as well as smartphones/tablets, should generally be mounted on tripods. However, mounting a camera on a tripod limits the size of the space and the type of activity that you can record. While you can capture a domestic space (e.g., a room of a house) or a smaller built space (up to perhaps 0.5 acre or 2,000 m^2) using multiple mounted cameras at different angles, you cannot record people moving over landscape-sized space using this technique. If you are specifically interested in recording landscape-scale motion, an action camera may be the best choice (Hermes et al. 2023).

Sometimes people are also tempted to use action cameras to record domestic space -- on the scale of a single room of a house. The reason is typically that they need to record inside very small rooms where it is hard to position a tripod. These situations are challenging, but you can often resolve them by being creative with the placement of your tripod-mounted camera, as discussed in §3.9. This is a better choice than using an action camera because it allows you a wider choice of cameras. Also, using an action camera can introduce issues with the post-processing and archiving of the footage, so it is wise to avoid it when possible – at least until action camera formats become more standardized.

3.3 Factors that don't matter

Now we want to point out a factor you should **not** consider when selecting a camera. This is audio inputs and formats. As we discuss in more detail in the next few sections, there are fundamentally different requirements for making usable quality video vs. audio recordings. Because of these differences, it is usually best to record video and audio on separate devices.

This is good news because it means you can consider a wider range of cameras – for example, cameras that don't have XLR microphone inputs. On the other hand, if you use only one device, you are restricted just to cameras that meet the audio input and audio format requirements of your documentation project – i.e., have XLR inputs and are capable of recording WAV/uncompressed audio. There are very few cameras that meet these specs. Additionally, the main one that currently does meet them, the Zoom Q8, should be avoided for image quality reasons (§3.2.1).

Another factor that doesn't matter is the maximum resolution of the camera, as long as it's at least 1080p. Recording in higher resolutions than this, such as 4k, has very little marginal benefit for language research (Seyfeddinipur & Rao 2020). It also has an immense marginal cost because of the increased size of the recordings and the processing challenges that come along with size. Unless and until we can more easily handle files that exceed 8GB, it is probably best to record in 1080p.

3.4 Make and model recommendations

We are making specific make and model recommendations because the camera options on the market have been fairly stable for the last 10-12 years. Prices are from late 2024 in the US.

3.4.1 Budgets over 1000 USD

If your budget is in this range, you are using only one camera, and you can transport a kit weighing >3kg, the best option is a semi-professional camera such as the Canon XA series (currently ranging from $1500 for the XA50 to $2300 for the most recent XA70 model). Among cameras which a documentation project could reasonably buy, these models produce the best quality image, have the widest variety of audio input options, and include fail-safes for if the battery is depleted, card full, etc. You do not need the latest model in the XA series. Older models (e.g., XA30 vs. XA70) have almost all the same options and are available refurbished (see §3.4.3 on refurbished cameras).

If you have this budget and are using only one camera, but you cannot transport 3kg, the best choice is currently a prosumer camcorder such as the Panasonic HK series (currently $560-$800) or Sony Handycam series (currently $229-$948). These produce an excellent quality image, but will look worse at night or in low-lit environments than a semi-professional camera. They also do not have as many audio input options.

If you are using two cameras with this budget, you could buy one semi-professional camera (e.g., Canon XA series) and one prosumer camcorder (e.g., Sony Handycam). This setup strikes a balance between delivering high image quality (since you have one semi-professional device) and avoiding budget and logistical challenges (since you don't need to pay for and transport two semi-professional cameras). As a slightly cheaper and lighter option, you could alternatively buy two prosumer camcorders. Or, if you have a large equipment budget and minimal need to control the size and weight of your kit, you could buy two semi-professional cameras (e.g., two Canon XA60s).

3.4.2 Budgets between 500 and 1000 USD

At this budget, you should purchase one prosumer camcorder, such as a Panasonic HK or Sony Handycam model. Buying two cameras is probably not feasible. If your project requires recording from multiple angles, your best option is to use one camcorder plus a tripod-mounted mobile device. If you can use a computer in your recording setting, you could also consider running a camcorder recording to its own card, plus a webcam recording to a laptop (see below).

3.4.3 Budgets under 500 USD

Projects with this budget should be able to purchase a prosumer camcorder on sale or second-hand. Buying second-hand/refurbished equipment from reputable sellers is a reasonable choice if your budget is too small for a new camcorder. However, if you are buying older equipment, you should confirm that you will be able to replace the batteries, since camera batteries are semi-consumable (§3.5) and can vary between models of the same make. If buying second-hand is not possible, you can consider recording via a mobile device, or looking for a way to share equipment with another team. It's best not to record with an extremely low-budget camera, including the Zoom Q8, unless you have absolutely no alternative, including a smartphone.

Another option for small budgets is using a webcam. These are very affordable (e.g., the EMEET C950 webcam is $25 and capable of recording in 1080p). In our experience, not many people in language documentation currently use webcams that are marketed for vlogging or recording/streaming with a computer. However, you could consider this style of camera if (1) it can be tripod-mounted, (2) it can record a wide shot, and (3) you are always able to use a laptop while recording, or the device supports recording to media such as an SD card (not only to a separate device or livestream).

3.4.4 Further sources of make and model advice

As a general rule, people that work in video production or do it as a hobby are a good resource for asking about specific camera options. However, be crystal clear about your technical specifications and budget when asking for make and model advice. People will tend to assume that your recording environment mirrors theirs—for instance, that you can drive to the location where you record. Make your constraints explicit; you might say something like "I'm not trying to optimize for weight, but I need to be able to walk 1km while carrying all of the recording kit."

3.5 Peripherals

Besides your camera, you will also need peripherals. The essential peripherals are as follows.

SD cards. Some cameras use microSD, others use the same size SD as audio recorders. You will probably want a larger size than for your audio recorder—64GB or 128GB is good for video, but for audio, typically sizes up to 64GB are used. Other technical specifications will also be different; check the specifications on your video camera. You do not need SD cards that are advertised as "rugged," because as discussed below, you will be carrying the camera and cards in a protective case.

Batteries. Camera batteries are not all created equal. Batteries that are manufactured by the same manufacturer as the camera often work better. However, generic batteries made by a brand such as Watson are also serviceable. You need two batteries per camera if recording less than one hour per camera per day: one in active use and one backup. You will need 1-2 more batteries per camera if you record more than one hour a day. Having more than four batteries is only necessary if they last less than one hour each or you have limited access to electricity.

On the topic of electricity, two to three hours of access to mains electrical power or generator power per day is more than sufficient to charge a set of camera batteries (in parallel). If you have less access to electricity than this, for example if you are using a solar panel and battery, you may want to consider either extra camera batteries or additional solar wattage. When costing these alternatives, remember that camera batteries lose their ability to hold charge over time and may need replacement in as little as 4-5 years.

Battery chargers. These may not be necessary if you can charge your battery from the camera. However, you should always have a backup means of camera battery charging, such as another camera or a mains charger.

Lens protection. You should protect your camera lens with a neutral density filter or UV filter while in use, and with a lens cap when it's not in use.

Tripod and tripod bag. Your tripod holds the camera while filming, and your tripod bag holds the tripod for transport. Be sure that your tripod is not too heavy or bulky to transport in your field conditions. Manufacturers market bigger tripods as being more stable, but we are not convinced these models are actually more stable in field conditions (e.g., on the ground or an uneven floor).

A protective/padded carrying case. You can use a specialty camera case or backpack, but these can attract unwanted attention, for example in airports and urban areas. An alternative that is not as noticeable is to pack the camera inside a sturdy plastic case (such as an ultra-heavy duty Tupperware) and pad it with material inside the case, such as silica gel packs, and outside the case, such as clothing. This is not as secure as a camera case because dropping the backpack or getting the backpack wet is more likely to damage the camera, but it does not make people as interested in you.

3.6 Testing and checklisting

To avoid problems with your video kit, thoroughly test it before the first time you use it and before each time you travel with it. It's especially important to test the batteries before each time you travel with the kit, as camera batteries lose their ability to hold charge over time.

After testing your kit, create a checklist of the components and keep it inside your camera case. Review the checklist each time you pack and unpack the case, as this kit contains many small parts that can easily get left behind. Also, practice setting up and taking down your camera and peripherals until you can do it efficiently. Use a checklist for this process too.

3.7 Audio setup for single-source recording

We haven't said anything yet about your audio kit. This is because the components of your audio kit for video recordings will depend on whether you are using a single-source or multiple-source audio recording setup. If you are using a single source—i.e., recording your audio on a microphone connected to the camera—you will need to be aware of the following points and buy the following additional peripherals.

3.7.1 Competing needs of audio vs. video recording

Unless you are recording a close-up of someone's face, which you should almost never do in language documentation (except for visual phonetics), your video recordings will need a wide shot. That is, to include the participant's entire body and surroundings in the frame, you will need to place the camera at least 2-3 meters from them and not zoom in.

A wide shot is necessary for the video—you need this to keep all relevant information in the frame (§3.9). But if your microphone is attached to the camera, it's terrible for the audio, because the participants are now too distant from the mic and can't be heard. To deal with this conflict while remaining single-source, you have two possible solutions.

3.7.2 Cardioid microphone and mic stand

First, you can connect the camera to a freestanding mic mounted on a mic stand, ideally one with a boom pole, and bring that closer to the speakers. It should probably be a cardioid microphone, since using a shotgun microphone (which is highly directional) will lead to large differences in volume if the participants move around. If you choose this option, you will need to buy a quality cardioid microphone—one example is the Rode NT4—and a mic stand.

The cardioid-and-mic stand setup can produce excellent audio of multiparty interactions, especially in situations with relatively little background noise. However, it is challenging to use because mic stands with booms are very bulky and heavy. They are made of steel and even the smallest ones are about a meter long when collapsed. If working in a location where you cannot buy mic stands locally, you will need to pack and check an additional bag when traveling to your destination solely for the mic stand. Ministands are more portable but vastly limit the placement of the microphone.

There are also other practical problems with using a mic stand. For instance, when you have to drape microphone cables a significant distance between the mic stand and the video camera, it adds complexity and time to setting up the shot. Furthermore, if people walk around the scene and bump into the mic stand or cable, this setup also creates trip hazards and a risk of the cables being pulled out, which can break the cable's XLR connector.

3.7.3 Wireless microphone system

As an alternative single-source method, you can use a wireless microphone system. This consists of 1-2 microphones proper plus a device which attaches to the audio input on your video camera. The Rode Wireless Go 2 is an example of a wireless microphone system which could be suitable for fieldwork. However, as mentioned in §3.1, this system requires connecting the wireless mic to the transmitter to the camera, which introduces two additional failure points.

3.8 Audio setup for multiple-source recording

Rather than buy a wireless microphone system or a cardioid mic and stand, the much simpler and cheaper option is to use a separate audio source (or sources). In this setup, you simply run an audio recorder near the participants at the same time as you run the video far from them (which you may already be using if you are doing audio documentation). Start one recorder, then the other, then clap in front of the camera to create a reference sound and image—you will later use this to sync the recordings.

If your participants are stationary during the recording, you can use whatever audio recording methods you'd use for an audio-only session, such as a Zoom recorder and lavalier microphones, or a Zoom recorder connected to a headset microphone for acoustic phonetics.

If your participants are moving during the recording, you'll need either (a) a wireless microphone system (§3.7.3) or (b) to record the audio track on devices the participants can wear. The first author records many scenes with 3+ participants that are all mobile; she uses an Olympus VP10 device (integrated recorder and microphone) on each participant. This device attaches to the participant's lapel, runs on battery power, and records to an internal hard disk. It is not a "wireless microphone system", but it is wireless, which means that the participants can walk around without the risk of pulling out cables. The Olympus VP10/VP20 and similar devices are comparable in price to wireless microphone systems, at approximately $100 per mic, but have fewer failure points. They are marketed for "business" rather than music recording use, so if selecting a device like this, be sure it can record in WAV format and an acceptable sampling rate.

3.9 Shot setup

So far all of this discussion has been about selecting and assembling your camera kit. What about once the session begins? You will need to select a good location for your camera(s), decide how to light the scene, and set up the shot on the camera (ie, decide whether to zoom in/out). Because other literature covers these considerations (Enfield 2013; Dingemanse 2011; Seyfeddinipur & Rau 2020; Pentangelo 2020; Ashmore 2008; Chrysanthi et al. 2016), our discussion here will be relatively short.

3.9.1 Camera location and lighting

Cameras should be mounted on tripods to prevent camera shake. If you have no choice but to film handheld, use something in the environment (or your own body) to stabilize the hand holding the camera. If filming in a small domestic space where it is hard to accommodate the tripod, consider taking the camera outside the space. You can position it looking in through the door or looking in through an open window. In a house with an elevated floor and no walls, as in parts of Amazonia and Southeast Asia, you can also put the tripod on the ground outside the house, raise the tripod height to its highest setting, and tilt the tripod head up to film people sitting on the elevated floor.

The main light source should be kept behind the camera as much as possible. This is necessary both to light the scene adequately and avoid glare from backlighting. Electric lights are good light sources, but you can still get adequate lighting without them, as long as you film in daylight. Windows and doors also make good light sources to place the camera near. In order to add light to the scene or reduce backlighting, you can ask participants to open/close windows and doors, and turn electric lights on/off if they have them. Candles and flashlights are not very effective means of lighting, but they're better than nothing if you need to film in a dark room without accessible windows or electric lights. Skin color can also affect lighting needs and camera palettes are optimized for lighter skin tones (Lewis 2019).

3.9.2 Use a single, wide shot

As Enfield (2008) and Seyfeddinipur and Rau (2020) recommend, you should avoid panning and zooming the camera as much as possible. Instead, set up one, wide shot and keep it the same. That is, only pan or move the camera if the participants are moving out of the frame.

There are multiple reasons for the recommendation of a wide, stable shot. One is theoretical: language documentation footage is meant to be a general-purpose record that is suitable for many types of future analysis. In contrast, panning and zooming are highly interpretive moves that focus on just one part of the participant's body (or other things in the environment), to the exclusion of the rest of the scene. Panning the camera or taking close-ups are selective in the same way that a researcher's notes on the visual context of an audio-only session are selective; they defeat most of the point of making video recordings.[14]

Another reason to use a wide, stable shot is that—most likely—you are not a professional camera operator. This means you will not be able to pan the camera as fast as the participants move. Therefore, panning and zooming in order to keep people in the frame are not likely to be successful. If participants move only occasionally and you pan in order to keep them in frame as they—for example—cross the room, you're more likely to succeed. Panning a lot is also distracting to watch later on.

To check the width of your shot, you can ask each participant to make a "snow angel" motion or a large pointing gesture after you set up the camera, but before you start recording. If you can see the entire "snow angel" and plenty of additional space on the camera's monitoring screen, your shot is probably wide enough.

If you can't see everyone's "snow angel" at the same time on a single camera, or if your shot passes the snow angel test but participants regularly end up out of frame due to movement, it is a sign that you need multiple cameras. When filming with multiple cameras, it's usually best to position them at either 90º angles or 180º angles to each other. Positioning two cameras at an 180º angle is helpful when recording two people who are interacting face to face with only a small space between them; each camera records one participant and their background. Positioning two cameras at 90º is helpful when recording people who are interacting with significant space between them: one camera records a focal participant (or group of participants or location), and the other camera records what is happening in the middle space.

Even if filming people who tend to sit on the floor, do not point the camera directly at the floor. The participants will get cut off if they stand up.

4 Post-processing

Audio files and video files require very different degrees and types of post-processing (that is, dealing with the recordings after they are made). Waveform audio files are ready to use from the moment you take them off the SD card. Video files, on the other hand, need to be processed in order to be used. This processing requires a number of steps: locating the files, concatenating clips into usable recordings, and converting the files between formats (transcoding) for archiving and annotation purposes. If the files are large, you may also need to reduce the file size before annotating or sharing (although not before archiving). This is called downsampling.

[14] If such shots end up being very desirable for publications, it's always possible to edit clips of a video to focus on a particular portion of the shot.

4.1 Locating your video files

Most video cameras do not format their SD cards in the same way as audio recorders. In particular, file browsers usually display the directories on a video SD card as packages (or hidden folders) instead of directories. Because of this, the easiest way to navigate to the recordings is using a complete path instead of interactively clicking into the directories. For many cameras, the path to the recording files from the top-level directory is something like `` `PRIVATE/AVCHD/BDMV/CLIPS` ``, but follow instructions specific to your camera.

For most cameras, the clips on the SD card will be in MTS format. Not all media players support this format. Make sure that you have a media player on your computer that does support MTS, such as VLC[15], so that you can play clips from your SD card before transcoding. Additionally, ELAN does not support MTS, which is one of the reasons you will need to transcode the files before working with them further.

4.2 Choosing a tool for video post-processing

Concatenating and transcoding video files requires a video editing software tool. FFmpeg[16] (Tomar 2006) is the best choice for most users, for several reasons: it has a large user community, it is free and open source, it allows total user control of all presets, and you can easily call it in bash scripts that you run from the command line. There are many user guides demonstrating how to install FFmpeg.[17] You can follow any of these guides. FFmpeg is general-purpose video editing software, not disciplinary software—it's more like Audacity than like Praat. Therefore, you don't need to stick only to FFmpeg support resources created by linguists.

Other consumer-oriented video editing software may come with your camera (e.g., Sony PlayMemories) or your computer (e.g., iMovie or QuickTime). This software is designed for creating slideshows, editing home movies to be visually appealing, adding frames, and so on. You might want to use it to create certain community-facing end products of the documentation from your recordings, but it is not what you need for concatenating and the next step, transcoding.

You can access FFmpeg via the command line or via a user-created GUI (e.g., https://qwinff.github.io/). We suggest accessing it via the command line, as most support resources for FFmpeg assume you are using the command line. Also, accessing FFmpeg via the command line will position you well to write shell scripts to automate FFmpeg processes later in your workflow, which will save a great deal of time. To access the command line, Mac

[15] https://www.videolan.org/vlc/

[16] https://www.ffmpeg.org/

[17] Such as the ELAR user guide at http://hdl.handle.net/2196/65788695-878d-44f9-a022-d9b54a83a441 or the generic Windows user guide at https://www.wikihow.com/Install-FFmpeg-on-Windows.

computers come with a Terminal program. For Windows, you are better off installing a Miniconda environment which has a Powershell.

4.3 Concatenating video files

Some video cameras split recordings above a certain file size into clips on the SD card. The maximum size of a clip depends on the technical specifications of the SD card and camera. The primary cameras which the first author has used, the Canon XA30 and Sony PJR540, create clips of approximately 4GB and 2GB respectively. With her presets, 4GB is about 34 minutes recording time and 2GB is about 17 minutes, meaning that an hour-long video recording will always contain multiple clips.

Before you transcode or annotate your video files, you'll want to concatenate the clips from each session into a single file. To concatenate your clips, you can use the concat command in FFmpeg. The sample command below allows you to concatenate the two files `session1_clip1of2.MTS` and `session1_clip2of2.MTS` without changing any of their video or audio encoding. ELAR also offers guidance on concatenating clips using FFmpeg.

```
`ffmpeg -i concat:session1_clip1of2.MTS|session1_clip2of2.MTS -c copy
session1_concatenated_file.MTS`
```

4.4 Transcoding video files

As mentioned in 'Locating video files', many cameras save files in MTS format. Other common camera formats include MOV (for the Zoom Q8) and MP4 (for Canon XA60+). MOV and MP4 are playable in ELAN for annotation, but MTS is not. Additionally, MOV and MTS are less widely accepted file formats for archiving than MP4, though they are allowed by PARADISEC.

Given this situation, if your camera records to MTS, you are best off transcoding to MP4 for both annotation and archiving. If it records to MOV, you do not need to transcode for annotation but may need to transcode for archiving, depending on your achive's requirements. If your camera records to MP4, you likely don't need to transcode for archiving purposes. You may still need to downsample MOV and MP4 video (reduce its size) for sharing or annotating, but this is a separate process from transcoding (changing the file format) and is covered in the next section.

You can transcode a file from MTS to MP4 without changing any of the encodings using FFmpeg. If you need to concatenate, do that first, then transcode the concatenated file. For an input file named `concatenated_file.MTS` and an output file named `transcoded_file.mp4`, the FFmpeg command is:

```
ffmpeg -i concatenated_file.MTS -y -acodec copy -vcodec copy transcoded_file.mp4
```

Once you have transcoded the concatenated file, you can delete both it and the unconcatenated items, since it can always be recreated from the MTS files. Practices differ on whether to retain the MTS files as well as the transcoded MP4 file after transcoding. If your archive requires you to upload both MTS and MP4, you of course need to retain the MTS files. Transcoding is usually not completely lossless, but compresses aspects of the visual signal which don't affect human viewing.[18] That is, it is visually lossless while still providing substantial compression.

Transcoding is the final step that you need to complete before archiving, which is discussed in the next section.

4.5 Downsampling

You can, and should, send your losslessly transcoded MP4 files to the archive. However, you probably won't be able to send these files to community members, collaborators, and so on for viewing, because they will be much too big for digital sharing. You also probably won't be able to annotate your losslessly transcoded videos in ELAN, since the program's performance declines quickly when you have video files linked that exceed ~2GB.

Because of these issues, you'll want to reduce the file size for sharing and annotation. While totally lossless compression of video isn't possible, you can often decrease the resolution of a HD video file without any real impact to its utility for annotation. For example, you can still reliably annotate eye gaze and other fine-grained visible behavior in a file where the original resolution of 1080p has been decreased to 720p. If a 720p video is still too large for sharing, even 480p can be acceptable quality to use for transcription.[19]

If your files need to be concatenated, it's probably most efficient to create the downsampled file directly from the intermediate MTS file produced by concatenation. If there's no concatenation, you can use either the losslessly transcoded MP4 copy or the format written by the camera (e.g., MTS) as the input to downsampling. The following FFmpeg command downsamples a MTS file `input.MTS` to an output named `output_720p.MTS` with 720p resolution and AAC audio:

```
ffmpeg -nostdin -n -i input.MTS -map 0 -c:v libx264 -preset ultrafast -crf 23 -s 1280x720 -
pix_fmt yuv420p -coder vlc -refs 1 -c:a aac -ab 256k -ac 2 output_720p.mp4
```

[18] https://trac.ffmpeg.org/wiki/Encode/H.264

[19] You can also reduce the frame rate or bitrate. However, we don't recommend reducing the frame rate (the number of times per second the image is updated). The most common frame rate, 29.94 fps, gives an update of the image approximately every 33 ms, which is close enough to align with most speech events. Reducing the frame rate below 25 fps will make it much harder to align physical gestures with the speech stream.

4.6 When to concatenate, transcode and downsample

During fieldwork, you should concatenate, transcode, and downsample on a daily basis as part of your backup procedure. These processes can and should be scripted to reduce their demand on your time. Refer to the Appendix or ELAR's FFmpeg user guide[20] for sample scripts.

Concatenating, transcoding, and downsampling all take significant processing time and electricity. Because of the time demand of these processes, it's best to run your post-processing scripts overnight or at another time when you can leave your computer on for several hours. This isn't possible in all field situations, especially if you don't have mains electricity. In these settings it's probably best to only post-process the files that you plan to annotate during the trip and leave the rest for when you return.

4.7 Aligning video and audio tracks

If you're using a multiple-source recording setup, you'll need to synchronize the video and audio tracks. This requires a reference sound that is hearable on the audio and also has a visual signal.

The classic reference sound is a clap. Just after you turn on the last recording device in the setup, clap your hands directly in front of the camera and within range of the audio devices. However, a single clap can be easy to confuse with other transients on the audio recordings, especially if there is a lot of background noise. To prevent this, you can clap twice in sequence or clap and then say a constant word/phrase. If you forget to clap, you can use another loud transient sound that's present on all of the recordings as a reference sound, for instance a noisy stop release. It's crucial that the reference sound be a transient with a clear onset, not a sonorant or another sound with a long decay.

To align the video and audio tracks, you can use either Synchronization Mode in ELAN or the Linked Files pane. Synchronization Mode seems to be the most popular option and is described in the ELAN manual. However, it supports a maximum of four media sources. If you have more than four media files associated with one ELAN file, you need to synchronize the media using the Linked Files pane instead. To do this, go to Segmentation Mode or Annotation Mode. Identify the time of the reference sound in each file (for example, by viewing the video or by locating the time of the reference sound on the waveform), then right-click on the waveform (for audio) or the video player (for video) and select 'Copy Recording Time without offset'. Then, navigate to Edit > Linked Files and set the offset for the relevant file to the copied value. Repeat this for each file.

[20] http://hdl.handle.net/2196/c1c22ff5-aeb5-4b78-bc34-183978eff708

If you want to play back your video outside of ELAN and the audio is on a separate source, you can use FFmpeg to replace the camera audio with the audio from the external source, using a command like the following:

```
ffmpeg -i my_video.mp4 -codec copy -an my_video_no_audio.mp4
ffmpeg -i my_video_no_audio.mp4 -i my_audio.wav -shortest -vcodec copy -acodec aac
my_video_new_audio.mp4
```

If you choose to replace the camera audio, you should be careful to check that there is no additional sound on the camera's audio track which you might need to use. For example, you probably shouldn't completely replace the camera audio with the recorder audio if someone is standing next to the camera and speaking to the participant. As an alternative, you can add the external audio as an additional audio track.

During transcription, you can switch back and forth between audio sources using the Controls panel in Segmentation or Annotation Mode of ELAN. This can be necessary in, for example, the situation described above where some speech is most hearable on the recorder audio and other speech is on the camera audio. However, switching audio sources consumes a lot of time, because it requires you to switch from Transcription to Segmentation/Annotation Mode and select a new source in Controls each time you switch (and there is no keyboard shortcut for either of these processes). Plan your transcription workflow to minimize source switching if possible.[21]

4.8 Adding subtitles

Depending on how you share your recordings, you may want to add subtitles or captions to the video. Transcriptions or translations will increase the accessibility of your recordings. There are several different caption formats, but they are all basically just text files with timestamp information. SRT is the most widely used format. If your subtitles start life in an ELAN file, you can export to SRT using the File > Export As… > Subtitles Text command.[22] If your subtitles start life in another time-aligned format, you can almost certainly make that into an SRT too. Once you have the SRT file, it can be used in many other applications, including playing in VLC or uploading to YouTube.

In order to embed video with subtitles in other programs (such as PowerPoint or PDFs), however, you will need to "burn" the captions into the video. When subtitles are burned in, they are part of the video stream rather than in a separate track, so they will be present whenever the

[21] For example, create intervals on tiers visually using different tracks and then transcribe them one at a time.

[22] Keep in mind that tiers aligned in ELAN for analysis might not always be directly usable as captions for sharing. For example, annotations are usually precisely coterminous with the utterance to be annotated, whereas captions usually have a time buffer.

video is played. FFmpeg is the easiest program to use to burn captions into a file. To do this, convert the SRT file to ASS format in FFmpeg, then burn in the ASS subtitles using the following commands:

```
ffmpeg -i my_subtitles.srt my_subtitles.ass
ffmpeg -i my_video.mp4 -vf ass=my_subtitles.ass my_video_burned_subtitles.mp4
```

Keep in mind that the burn-in will take a long time (at least twice the recording time), because you are overlaying new pixels with the subtitles on top of every frame of the video. Because it changes the actual information in the video frame, once subtitles are added via burn-in, they can't be removed – so don't do it on your original recordings!

4.9 YouTube

Finally, we should mention the role of YouTube in video documentation. YouTube is a very popular video sharing platform. Because it is a sharing platform, not an archive, it should definitely not be the sole, or even primary, long-term home of your data. YouTube provides a straightforward way to share materials for streaming (that is, viewing-only). It isn't a solution for sharing for annotation, because a) it compresses files for upload and optimal streaming, and b) downloading files from the site is both inconvenient and in violation of the terms of service. You should also be aware of YouTube's user agreements, which involve granting YouTube a perpetual license to the data you upload. Burke et al. (2022) discuss examples of alternative ways to share video with community members.

To upload a video to YouTube, you'll need an account. Use the "YouTube Studio" option, which will take you to a dashboard to "upload videos". There will be questions about them, and then YouTube will upload the videos and ingest them into the site. While they are uploading, you can provide metadata to allow people to find the recordings (or keep them unlisted if you only want people with the link to see them). You can also add captions. YouTube will auto-caption videos if you specify the language (which you can then edit), or you can upload your own.

5 Archiving

After your recordings have been made and processed, they will need to be archived, along with other work in the course of the documentation project. This section covers considerations related to the archiving of video documentation specifically. There are many general archiving methods guides available (including Johnson 2004; Andreassen 2022; Conathan 2011). Video documentation brings with it all the usual data management considerations around file structure, naming conventions, metadata and the like (Bowern 2015). Two points in particular stand out: the additional storage needs associated with large files, and additional issues of ethics, particularly anonymity. We address both in turn.

5.1 Storage

Because video is space-intensive, long-term storage (both local and in a digital archive) is a more significant issue if using extensive video documentation compared to audio alone. Because some archives must pay for storage, they have limits on what they can accept. Some archives recoup some of their costs with fees to depositors. If you are responsible for fees to the archive, make sure to inform them you are shooting video and give an estimate of total GB. As a rule of thumb, 1 hour mp4 video at 1080p resolution, a frame rate of 30, and a video bitrate of 15 is 6.7GB. In comparison, 1 hour of audio recorded at 16 bit depth, 44,100 Hz and stored uncompressed is 635MB. Archives that accept video include ELAR, CLA, and PARADISEC (PARADISEC does not specify a target bitrate).

One can sometimes use university repositories (e.g., dataverses)[23] for archiving, but they may have their own requirements, limits, constraints on who may be a depositor and promises for long-term preservation. It is better to archive materials with a language archive if possible. Archiving in multiple different places also leads to the fragmentation of linguistic collections which impedes discoverability (cf. Babinski et al. 2021, Yi et al. 2021 for more discussion). That said, it is better to preserve materials than to rely on luck, so if you cannot archive all your video with the rest of your collection, a dataverse or repository like Zenodo may be the best of many dispreferred options.

The storage of large (>500 GB) collections of data for local use also requires some thought, and the issues become more acute as collections increase in size. As universities move away from unlimited cloud storage (e.g., given that Google recently removed its education unlimited storage plans), this problem is likely to increase. The shift to tighter storage limits and deletion of unused accounts also raises longevity issues and makes it all the more important to be careful about backing up data appropriately. Material can be stored on solid-state external drives, but that is not a long-term solution without a plan for the long-term management of the data. Working copies of data on external hard drives (e.g., solid state drives that hold several TB) is the current best solution, a tradeoff between portability and price. They can be mounted on a networked computer to create a local shared drive. You may need to store both archival quality and downsampled working copies, at least temporarily (§4.5).[24]

[23] Repositories are sites which provide long-term stable access to materials of many different types. They are unlike archives in that they are typically minimally curated, they are depositor-driven (that is, depositors usually control decisions on what content to upload), and they may or may not provide undertakings for long-term preservation. That is, they provide a location for serving content but do none of the curatorship and stewardship tasks that archives commit to.

[24] Another option is to purchase a NAS system, which is both networked and can create backups (Synology DiskStations are a reasonable price point and are well supported and flexible). However, for use beyond the local

5.2 Data formats

The standard archival format for video has been MP4 for many years now (cf. Seyfeddinipur & Rau 2020). ELAR and PARADISEC offer more detailed advice on audio and video specifications – for example, PARADISEC also specifies a frame rate. This is another reason to plan archiving *before* starting recording; always have a concrete archiving plan in place and be clear about the requirements of the archive you plan to work with in advance. In our experience, however, archives may not be able to give a great deal more information about file specifications than "mp4", so we recommend following ELAR's or PARADISEC's more detailed information (e.g., https://paradisec-archive.github.io/PARADISEC_workflows/04_standard_formats.html)

You should make sure that you archive all relevant files associated with the project. For video, this might mean not just the transcription files, but also settings files that provide time offset and linked file information.

5.3 Ethical issues related to archiving

Linguistic documentary material is partly a research resource, partly general testament to time, place, and people, and partly ongoing community resource, as the literature on language collections make clear. The more naturalistic the recordings, the harder the line is to draw between personal, private, or public information. Some general considerations that arise when making video data public were given in §2.3 above. In this section, we focus on the impacts of ethics on archiving and data processing.[25] Different countries have different rules, and data protection regimes and you will need to comply with both your home country's rules and those of the country you record in (if they are different).

First, just as with other types of fieldwork recording, you will need to decide what material should end up in the archive. The archive does not necessarily need every single file you recorded, just as they don't necessarily need every single envelope you wrote something on the back of. Part of doing language documentation is balancing rich recording and deep description with other needs, including the privacy of individuals, the purpose of the recordings, and likely future use. Examples of things that you wouldn't need to keep (and certainly not archive) due to future use include recordings of false starts or failure to turn off all recordings at the end of a session. If a recorder kept running, you should trim the file before annotation. For privacy reasons you might not want to archive personal or community gossip, or personal discussions about other individuals. Research participants should be the guide to these content decisions. As

area network they require an understanding of firewalls and web security which may not be something you wish to acquire. That said, they can be used as LANs for local backup and file access.

[25] See https://kelseycneely.com/files/Neely-Preparing-data-for-open-access-May2020.pdf for specific workflows and examples for many of the points in this section related to anonymizing video data.

Hanke & Fenlon (2022) note, ideally you discuss such items with participants, either around the time of the session or by rewatching the video together. Such items can be flagged during the recording, or in a separate tier during transcription and annotation.

Secondly, your language documentation project likely went through university ethics approval, and you obtained informed consent from research participants before you started. What did you agree to? What did your IRB require? What did your research participants want? If they didn't agree to having videos of themselves placed in archives, you can't archive those video recordings (or put them in other repositories either). Did you promise full anonymity and/or confidentiality? If you promised anonymity, you can't archive video[26], since 'anonymity' means that the participants' details are not recoverable, and images are intrinsically identifiable. Promises of confidentiality should be accompanied by more detailed information and negotiation about acceptable uses for video and audio recordings and how they should be associated with individual participant identities. To be clear: you should not promise anonymity or confidentiality if you cannot follow through, and promising this will mean you can't archive a lot of your data. If you can't archive it (and if you can), you need to be clear about what should happen to it.

It's helpful to class materials into four categories, which will help you figure out what needs to happen to set type of recording: 1) public (no restrictions to associated recordings and metadata, placed in accessible archive); 2) redacted (public but with some information removed for archiving purposes; 3) restricted (accessible only to specific individuals or groups, with or without redaction); this includes material that should be preserved (and can be preserved without legal issues) but might not be appropriate for general availability; and 4) not-archived (e.g., material that cannot be on a server for legal reasons, things that shouldn't have been recorded in the first place, or errors in recording that could be edited out).

These actions are a necessary but time-consuming part of curating your recordings. Allow time for it. Redaction may involve audio, video, or both. For example, some collections redact the names of individuals, since associating specific people with the language project may endanger their safety by identifying them as a member of that ethnic group. In such cases, one can either search transcripts for specific words or use San et al.'s (2021) approach to keyword identification in speech to speech search to find them (however, this may not be fully reliable). The types of information that may need to be redacted will vary. Video redaction also varies, but some classes of material for video redaction include bystanders who didn't give their permission to be filmed and babies and children who aren't wearing clothes (posting videos of naked children online is an offense in many jurisdictions, even if it isn't for pornographic purposes, and can lead to storage accounts being deleted with no opportunity to restore). If bystanders who enter the shot are potentially recognizable, you need to determine if these bystanders consent to having their

[26] You probably can't archive audio either, especially in small communities.

image archived. If not, the bystanders' images should be blurred before the data is archived. To blur parts of a video, you can use Adobe Premiere (part of the Creative Cloud Suite); the resource by Kelsey Neely cited in note 24 offers detailed instructions and screenshots for how to make these changes effectively.

Remember that redaction may remove information relevant to the research. Blurring faces, for example, will remove information about gaze. That should lead you to think about the integrity of the "collection" vs. a research "corpus" for replicability of research and subsequent use. That is, linguists often think of their research materials as a single set of items, but in practice they are often better thought of as a set of interlinked research resources, some of which may stand alone and some of which do not.

In some cases it may be preferable to archive only the audio of a session (or address situations that might prevent archiving during the session itself). For example, it is not a problem to archive an audio recording of someone holding an undiapered baby, but processing the video of the same session for archiving will require a very time-consuming blurring stage. In such a case you could not record the video; you could provide a diaper; or you could arrange, if appropriate, for the child to be out of the video frame.

6 Summary of Recommendations

Thus in summary, we argue in this tutorial that video is an important part of language documentation. We provide concrete recommendations for equipment and workflow and summarize the additional archival considerations that arise when making visual recordings. In this section, we briefly summarize each of the recommendations from §§2-5 above.

Record video where possible, even for documentation which prioritises speech as a linguistic modality. The utility of the visual information outweighs the cost of additional processing and storage. Visual information links speech to the surrounding context and provides richer documentation than audio alone (or audio and written notes).

Before selecting a camera, peripherals, and data processing dataflow, consider a) the type of events you will record; b) whether you will record with a separate audio device or with audio linked to the camera; and c) how many cameras you will need and where they will be placed. For the choice of camera, consider price, size, and mobility, but not the audio quality (if you can use a separate audio recorder) or the maximum recordable resolution. Use a cardioid microphone and stand, headset mic, or wireless mics with your camera(s). Make sure to have sufficient SD cards, batteries, and good means of charging them, along with other peripherals that protect the camera. Compile a checklist of all the equipment and use it when you set up pack up. Invest time in learning how to use the equipment, how to set it up and control its features. Practice setting up shots. Take into account the camera location, lighting, and what's in the visual frame. Practice making recordings that don't cut off participants or make their features difficult to see.

Before you record, install FFmpeg and choose a set of commands or scripts for postprocessing video. The three tasks before annotation and sharing are concatenation, transcoding, and compression. Do this as soon as you can: in the field if possible, but since the files take substantial time to process and should be done with a consistent power source, it may need to wait until your return. After taking the files from the SD card, transcode to MP4 or MOV format for preference. Archive lossless transcoded files, but downsample to 720p for annotation. Subsequent video processing may be necessary when sharing clips (e.g., burning in subtitles before including video material in articles or talk slides).

For annotation, align tracks made with separate audio and video in ELAN. Record a distinct audio-visual reference point (e.g., a clap) at the start of the recording to use as a reference point. Align in ELAN's synchronization mode or using the Linked Files pane.

Archive your recordings. Proceed as you would for creating archival collections for other types of documentary data. Be clear about the ethical implication of recording visual data, because participants are more identifiable and the data are richer. This richness is ideal for linguistics and language reclamation, but it also brings additional challenges. When archiving, make sure your practices align with your promises and the legal requirements for your institution, the communities you work with, and the countries with jurisdiction. Separate material into public, redacted, restricted, and private/non-archived categories, and redact information by blurring the video.

Last, look for a community of practice around video recording. As in all research, technical and ethics standards for working with video will change over time, and you can most easily learn about these changes from others doing similar work. Seek out other people who record video for language documentation—or in allied fields like visual anthropology and sign language linguistics—and bring your ideas and questions to them.

References

Andreassen, Helene N. 2022. Archiving Research Data. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.), *The Open Handbook of Linguistic Data Management*, 89–100. Cambridge: The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0011.

Ashmore, Louise. 2008. The role of digital video in language documentation. *Language Documentation and Description* 5. 77–102. https://doi.org/10.25894/ldd253.

Berez-Kroeker, Andrea L., Shirley Gabber & Aliya Slayton. 2023. Recent Advances in Technologies for Resource Creation and Mobilization in Language Documentation. *Annual Review of Linguistics* 9. 195–214. https://doi.org/10.1146/annurev-linguistics-

031220-120504.

Blythe, Joe, Rod Gardner, Ilana Mushin & Lesley Stirling. 2018. Tools of Engagement: Selecting a Next Speaker in Australian Aboriginal Multiparty Conversations. *Research on Language and Social Interaction* 51(2). 145–170. https://doi.org/10.1080/08351813.2018.1449441.

Bowern, Claire. 2015. *Linguistic fieldwork: A practical guide*. Dordrecht: Springer.

Burke, Mary, Oksana L. Zavalina, Shobhana L. Chelliah & Mark E. Phillips. 2022. User needs in language archives: Findings from interviews with language archive managers, depositors, and end-users. *Language Documentation & Conservation* 16. 1–24.

Chrysanthi, Angeliki, Åsa Berggren, Rosamund Davies, Graeme P. Earl & Jarrod Knibbe. 2016. The Camera "at the Trowel's Edge": Personal Video Recording in Archaeological Research. *Journal of Archaeological Method and Theory* 23(1). 238–270. https://doi.org/10.1007/s10816-015-9239-x.

Church, R. Breckinridge, Martha W. Alibali & Spencer D. Kelly. 2017. *Why Gesture?: How the hands function in speaking, thinking and communicating*. Amsterdam: John Benjamins.

Conathan, Lisa. 2011. Archiving and language documentation. In Julia Sallabank & Peter K. Austin (eds.), *The Cambridge Handbook of Endangered Languages*, 235–254. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511975981.012.

Cychosz, Margaret, Rachel Romeo, Melanie Soderstrom, Camila Scaff, Hillary Ganek, Alejandrina Cristia, Marisa Casillas, Kaya de Barbaro, Janet Y. Bang & Adriana Weisleder. 2020. Longform recordings of everyday life: Ethics for best practices. *Behavior Research Methods* 52(5). 1951–1969. https://doi.org/10/gng7wz.

D'Arcy, Alexandra & Emily M. Bender. 2023. Ethics in Linguistics. *Annual Review of Linguistics* 9. 49–69. https://doi.org/10.1146/annurev-linguistics-031120-015324.

Deal, Amy Rose. 2015. Reasoning about equivalence in semantic fieldwork. In M. Ryan Bochnak & Lisa Matthewson (eds.), *Methodologies in semantic fieldwork*, 157–174. Oxford: Oxford University Press.

Defina, Rebecca. 2016. Do Serial Verb Constructions Describe Single Events? A Study of Co-Speech Gestures in Avatime. *Language* 92(4). 890–910.

Degai, Tatiana, David Koester, Jonathan Bobaljik & Chikako Ono. 2023. The Siberian World. In Vladimir Davydov, Jeanne Ferguson & John Ziker (eds.), *Kŋaloz'a'n Ujeret'i'n Ŋetełkila'n—Keepers of the Native Hearth*. Abingdon: Routledge. https://doi.org/10.4324/9780429354663-5.

Dimmendaal, Gerrit J. 2010. Language description and "the new paradigm": What linguists may learn from ethnocinematographers. *Language Documentation & Conservation* 4. 152–158.

Dingemanse, Mark. 2011. *The meaning and use of ideophones in Siwu*. Nijmegen: Radboud University PhD thesis.

Enfield, Nicholas. 2013. Doing fieldwork on the body, language, and communication. In Cornelia Mueller, Alan Cienki, Ellen Fricke, Sylvia H. Ladewig, David McNeill & Sedinha Tessendorf (eds.), *Body -- Language -- Communication*, 974–981. Berlin: De Gruyter Mouton.

Engman, Mel M. & Mary Hermes. 2021. Land as Interlocutor: A Study of Ojibwe Learner Language in Interaction on and With Naturally Occurring 'Materials.' *The Modern Language Journal* 105(S1). 86–105. https://doi.org/10.1111/modl.12685.

Fenlon, Jordan, Adam Schembri, Trevor Johnston & Kearsy Cormier. 2015. Documentary and Corpus Approaches to Sign Language Research. In *Research Methods in Sign Language Studies*, 156–172. Hoboken: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118346013.ch10.

Goico, Sara Alida. 2019. *The Social Lives of Deaf Youth in Iquitos, Peru*. San Diego: University of California PhD thesis.

Hanke, Thomas & Jordan Fenlon. 2022. Creating corpora: data collection. In Jordan Fenlon & Julie A. Hochgesang (eds.), *Signed Language Corpora*, 18–45. Washington, DC: Gallaudet University Press. https://doi.org/10.2307/j.ctv2rcnfhc.

Hanks, William F. 2009. Fieldwork on deixis. *Journal of Pragmatics* 41(1). 10–24. https://doi.org/10.1016/j.pragma.2008.09.003.

Henke, Ryan & Andrea L. Berez-Kroeker. 2016. A Brief History of Archiving in Language Documentation, with an Annotated Bibliography. *Language Documentation & Conservation* 10. 411–457.

Hermes, Mary Rose, Mel M. Engman, Meixi & James McKenzie. 2023. Relationality and Ojibwemowin in Forest Walks: Learning from Multimodal Interaction about Land and Language. *Cognition and Instruction* 41(1). 1–31. https://doi.org/10.1080/07370008.2022.2059482.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). https://doi.org/10.1515/ling.1998.36.1.161.

Hochgesang, Julie. 2020. Sign Language Description: A Deaf Retrospective and Application of Best Practices from Language Documentation. Presentation. https://doi.org/10.6084/m9.figshare.13393427.v1.

Hochgesang, Julie A. & Jordan Fenlon (eds.). 2022. *Signed Language Corpora*. Gallaudet University Press. https://doi.org/10.2307/j.ctv2rcnfhc.

Hochgesang, Julie & Nick Palfreyman. 2022. Signed language corpora and the ethics of working with signed language communities. In Jordan Fenlon & Julie A. Hochgesang (eds.), *Signed Language Corpora*, 158–195. Washington, DC: Gallaudet University Press. https://doi.org/10.2307/j.ctv2rcnfhc.

Hou, Lynn Yong-Shi. 2016. "Making hands" : family sign languages in the San Juan Quiahije community. Austin: University of Texas dissertation.

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. In Peter Austin (ed.), *Language documentation and description*, vol. 2, 140–153. London: Hans Rausing Endangered Languages Project.

Ko, Edwin & Schuyler Laparle. 2022. Co-speech gesture and semantic fieldwork: A case study of aspectuals in Crow. *Semantic Fieldwork Methods* 4(3). 1–25. https://doi.org/10.14288/sfm.v4i1.195494.

Labov, William. 1972. *Sociolinguistic Patterns*. Oxford: Blackwell.

Lähdesmäki, Tuuli, Eerika Koskinen-Koivisto, Viktorija L.A. Čeginskas & Aino-Kaisa Koistinen. 2020. *Challenges and Solutions in Ethnographic Research: Ethnography with a Twist*. London: Routledge. https://doi.org/10.4324/9780429355608.

Leonard, Wesley Y. 2017. Producing language reclamation by decolonising 'language.' *Language Documentation and Description* 14. 15–36. https://doi.org/10.25894/LDD146.

Lewis, Sarah. 2019. The Racial Bias Built Into Photography. *The New York Times*, sec. Lens, April 25, 2019.

Mead, Margaret & Gregory Bateson. 1951. *Bathing babies in three cultures*.

http://www.colonialfilm.org.uk/node/4771.

Mitsuhara, Teruko Vida & Jan David Hauck. 2022. Video Ethnography: A Guide. In Sabina M. Perrino & Sonya E. Pritzker (eds.), *Research Methods in Linguistic Anthropology*, 223–259. London: Bloomsbury Academic.

Moriarty, Erin. 2020. Filmmaking in a linguistic ethnography of deaf tourist encounters. *Sign Language Studies* 20(4). 572–594. https://doi.org/10/gg96bn.

Nathan, David. 2009. The soundness of documentation: an epistemology for audio in documentary linguistics. Presentation. http://hdl.handle.net/10125/5103.

Nathan, David. 2010. Sound and unsound practices in documentary linguistics: towards an epistemology for audio. *Language Documentation and Description* 7. 262–284. https://doi.org/10.25894/ldd233.

Pentangelo, Joseph. 2020. *360° Video and Language Documentation: Towards a Corpus of Kanien'kéha (Mohawk)*. New York: City University of New York PhD thesis. https://www.proquest.com/docview/2459231889/abstract/F92D85A70B9B4FEAPQ/1

Rice, Keren. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics* 4(1–4). 123–155. https://doi.org/10/cgdbb6.

Schembri, Adam. 2010. Documenting sign languages. *Language Documentation and Description* 7. https://doi.org/10.25894/ldd228.

Seyfeddinipur, Mandana & Felix Rau. 2020. Keeping it real: Video data in language documentation and language archiving. *Language Documentation & Conservation* 14. 503–519.

Skilton, Amalia. 2019. *Spatial and nonspatial deixis in Cushillococha Ticuna*. Berkeley: University of California PhD thesis.

Tomar, Suramya. 2006. Converting video formats with FFmpeg. *Linux Journal* 146. 10.

Tsikewa, Adrienne. 2021. Reimagining the current praxis of field linguistics training: Decolonial considerations. *Language* 97(4). e293–e319. https://doi.org/10.1353/lan.2021.0072.

Turin, Mark, Claire Wheeler & Eleanor Wilkinson (eds.). 2011. *Oral Literature in the Digital Age*. Cambridge: OpenBook Publishers.

Woermann, Niklas. 2018. Focusing ethnography: theory and recommendations for effectively combining video and ethnographic research. *Journal of Marketing Management* 34(5–6). 459–483. https://doi.org/10.1080/0267257X.2018.1441174.

Zabulis, Xenophon, Nikolaos Partarakis, Valentina Bartalesi, Nicolo Pratelli, Carlo Meghini, Arnaud Dubois, Ines Moreno & Sotiris Manitsaris. 2024. Multimodal Dictionaries for Traditional Craft Education. *Multimodal Technologies and Interaction* 8(7). 1–48. https://doi.org/10.3390/mti8070063.