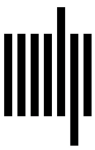
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

40 Managing Sign Language Video Data Collected from the Internet

Lynn Hou, Ryan Lepic, and Erin Wilkinson

1 Introduction

Research on spoken languages relies extensively on the use of written text. The speech signal is a continuous stream of acoustic information, and spoken language is usually accompanied by visible information such as facial expressions and co-speech gestures. However, linguists are comfortably accustomed to analyzing spoken language data using standardized text systems such as the International Phonetic Alphabet and the Leipzig Glossing Rules (but see, e.g., Pawley & Syder 1983 and Linell 2011 on the biases caused by over-reliance on written text conventions). Written language is also the primary object of study in the analysis of literary and digital texts. In contrast, while there have been a number of attempts to create orthographic systems for sign language users such as Sutton SignWriting, and transcription systems for sign language linguists such as the Hamburg Notation System, no system has yet been adopted as a suitable and generally accepted standard for textually representing sign language data (see Crasborn 2015). Sign languages are essentially unwritten, and this poses a considerable challenge for the representation, management, and accessibility of sign language data.[1]

Though sign language researchers have not reached general consensus on a standardized system for text-based representation of sign forms, research on sign languages has nevertheless progressed, through the common practice of representing signs with metalinguistic, meaning-based glosses. For American Sign Language (ASL), researchers typically use English glosses to represent manual signs and, when necessary, they superimpose these glosses with additional diacritics to represent facial expressions and body movements. Such meaning-based glossing is highly idiosyncratic and is fundamentally shaped by the researcher's analysis of the phenomenon at hand. Thus, readers often can only guess which signs are referred to and must imagine how glossed examples would be signed. This means that when reading scholarly publications about a given sign language, even competent users of the language generally do not have adequate access to the primary data being discussed. Researchers have adopted different strategies to address this issue of data accessibility. For example, Edward Klima and Ursula Bellugi's (1979) essential volume, *The Signs of Language*, popularized representing some signs with line drawings of human bodies in motion, and many of the illustrations that artist Frank Paul created for the volume, and others from around that time, continue to be widely circulated and re-used to this day.

However, as video recording and digital storage have become increasingly accessible and affordable, so too have they been increasingly embraced by signers and linguists alike to capture sign language data for future viewing.[2] One outcome of this use of digital video recording has been an expansion of sign language videos on the internet, particularly for ASL. Among other things, such videos open up new opportunities for researchers to work with naturalistic data from members of the ASL-signing community and to mitigate some of the problems of accessibility caused by the lack of standardized text-based systems for representing sign language data.

Accordingly, in this Data Management Use Case, we illustrate some of the practical considerations for working with digital video recordings of sign language data for language description purposes. We describe our efforts to analyze ASL signing on the internet, from data collection to sharing and citing, in a way that (we hope) can serve as a working guide for readers who may want to work with video data on the internet for the purposes of (sign) language description and documentation.

2 Working with digital video recordings of elicited (sign) language data

Here, we briefly describe some practical considerations for managing digital recordings of elicited sign language data. In many ways, managing sign language data is quite similar to managing spoken language data (see, for example, Holton, Leonard, & Pulsifer, chapter 4, this volume; Mattern, chapter 5, this volume; Kung, chapter 8, this volume; Hoey & Raymond, chapter 20, this volume; and Daniels & Daniels, chapter 26, this volume). One primary difference between spoken language research and sign language research is that, as described in section 1, sign language linguistics lacks a standardized transcription system for representing visual sign language data textually (see Palfreyman, chapter 21, this volume; Crasborn, chapter 39, this volume; and Hochgesang, chapter 30, this volume). As a result, all textual data and meta-data must be managed using the written form of another language, such as English.

High-definition digital video recordings are currently the preferred, if not the standard, means for recording and archiving sign language data. At a minimum, researchers will need at least one video camera, a tripod, a high-capacity memory card, a computer, video editing software, and an external hard drive or cloud service for storing large video files. Tight shots of multiple signers participating in signed discourse are best captured with multiple video cameras, including (at least) one camera for each signer. However, when working with multiple cameras, it also becomes necessary to synchronize all video recordings in the coding process, to simultaneously view both participants on the computer screen. The choice of video editing software may depend on editing requirements and technical knowledge. Investing in a large, reliable storage and data plan will save you from the heartache of losing your data.

Laboratories, filming studios, and other controlled environmental settings are very common for maximizing the quality of sign language data collected, especially when researchers are building a sign language corpus and filming signers with at least two cameras (e.g., Fenlon et al. 2015). Such settings allow researchers to construct an optimal layout for multiple participants and cameras (Perniss 2015). There are two main types of sign language data commonly distinguished in the field of sign language research, *elicited* and *naturalistic*. For collecting elicited data, researchers utilize a variety of elicitation tasks that involve visual stimuli such as pictures and video clips and written stimuli such as children's stories, semi-structured interviews for obtaining lexical and grammatical constructions, and elicited narratives. For collecting naturalistic data, there are at least two approaches. In sociolinguistic variation studies, researchers videotape groups of signers for one- to two-hour periods in public and/or social spaces such as deaf club events, schools, and conferences (Lucas, Bayley, & Valli 2001; McCaskill et al. 2011). In corpus studies, researchers may assign a pair of acquaintances or friends, or a pair of interlocutors who share similar demographic backgrounds such as age and/or region, and prompt them to engage in spontaneous conversation for half an hour in a filming studio (Fenlon et al. 2015). The idea is to reduce the effects of the "observer's paradox" (Labov 1972) by making some methodological accommodations with the intrusive nature of videotaping to obtain more naturalistic signing from the participants.

Linguistic research that involves primary language data collection from consultants typically falls under the umbrella of human subjects research, meaning that all data collection procedures should be approved by an Institutional Review Board (IRB) or a similar ethical committee. Fortunately, many descriptive linguistic projects are considered socio-behavioral (as opposed to medical) research, and because humans naturally use language every day, there are few risks associated with descriptive linguistic research, outside of boredom and fatigue for language consultants. When working with video data, however, the researcher also needs to be aware, and communicate to the ethical review board as well as to language consultants, that video data poses an inherent risk of loss of confidentiality and privacy for consultants. Typically, it is sufficient to anonymize data by assigning consultants pseudonyms (unless they request their real names to be used) in all data management and reporting, and to include questions on the consent form that ask consultants to determine whether their videos (1) can be viewed by other researchers, (2) can be shown at academic conferences, (3) can be published in academic papers, and, ideally, (4) can be archived in data repository sites for future research.

3 Working with digital video recordings of (sign) language data from the internet

In section 2 we outlined basic considerations for recording, managing, and sharing digital recordings of sign

language data for language description purposes. In this section we describe the additional advantages and issues that arise when analyzing sign language data from videos on the internet. Here we primarily consider ASL, the language that we have worked with in this capacity. As a globally dominant sign language, ASL is relatively well attested in signing videos on the internet. However, we expect that the suggestions we make here should be broadly applicable to any body of video data on the internet, for spoken and sign languages alike.

In essence, studying ASL signing on the internet involves searching for videos that have been produced by members of the ASL-signing community, creating a researcher copy of the videos, and analyzing them. In comparison to researcher-elicited video data, which we consider to include semi-naturalistic data obtained in controlled laboratory settings, what sets internet-based studies apart is that signers decide on their own to share their language in public internet spaces. Table 40.1 outlines the differences stemming from there, providing a general overview of the points made throughout this chapter.

3.1 Language sampling from internet data

Traditionally, research on ASL structure and use has targeted only a limited population of deaf signers who use one variety of ASL, and this pattern of recruitment follows from particular language ideologies and language attitudes that researchers hold about ASL and its users (Hill 2013). However, the ASL-signing community is heterogeneous, and the conservative approach that researchers take in recruiting language consultants has shaped how ASL varieties are represented to the wider research community. Acknowledging these facts leads to a bigger question: What does it mean to analyze ASL that is representative of the ASL-signing community?

Table 40.1
Comparison of ASL data types: Researcher-elicited data versus internet-based data

	Researcher-elicited	Internet-based
Interaction between participants and researchers		
Participant recruitment	Required	No recruitment needed
IRB (or ethical review board) protocol	Required	Depends on the institution and research question; if contacting individuals to collect information, then IRB is required
Profile of participant pool	Typically homogenous (e.g., educated, white, deaf of deaf, able-bodied) Individual recruitment	Potentially diverse; however, analyses (e.g., variationist studies) will depend on research goals and available data
Data authenticity		
Data elicitation	Planned elicitation tasks, including prompted naturalistic conversation	No elicitation involved
Observer's paradox	Although researchers may create naturalistic environments for consultants, knowing that video recordings will be shared publicly may affect language use	Contributors voluntarily post their videos in public forums (e.g., Facebook and Twitter), indicating that they acknowledge that others will view and even share their videos
Data sampling and sharing		
Genres	Typically limited variety	Typically wider variety
Language sampling	Given the nature of elicitation tasks, including naturalistic conversation in controlled settings, elicitation could generate adequate materials for very particular research questions	While internet-based ASL data may be more representative in terms of language usage, researchers may not easily find adequate materials to answer very particular research questions
Data distribution by researchers (for example to archives or repositories)	Not standard; researchers typically maintain closed databases	Any online videos that are available to one researcher are also available to *any* researcher, for as long as they remain public. However, it is still necessary for researchers to consider archiving videos

Historically, well-educated, white, deaf signers from signing deaf families have been considered representative users of a prestige variety of ASL. This prestige variety is what is represented in many scholarly publications on ASL. Some exceptions include research on sociolinguistic variation in ASL among Black and white deaf signers of varying socio-economic backgrounds (Lucas, Bayley, & Valli 2001; Lucas et al. 2001; Lucas & Bayley 2005) and the Black ASL Project, a study of a constellation of ASL varieties that emerged in segregated residential schools for Black deaf children and transmitted to subsequent generations of users in Black deaf families and communities (McCaskill et al. 2011). The Black deaf community is identified as "one of the underrepresented, underdocumented, and underreported populations in the literature on ASL and Deaf studies" (Hill & McCaskill 2016:62). Thus, it seems that the variety of ASL represented in most sign language research can be considered largely based on a specific demographic group, rather than as a representative sample of the ASL-signing population.

Fewer than 10% of the deaf signing population are born into signing families; these individuals are referred to as native signers or deaf (children) of deaf (parents). While there is no systematic and direct data available about ASL usage in the United States (Mitchell et al. 2006:307),[3] approximately 95% of American deaf children are born to hearing parents who do not know any sign language. Despite the heavy representation of native-signing ASL users in the literature, then, the majority of deaf, signing Americans do not acquire ASL from fluent signing caregivers as children (Mitchell & Karchmer 2004). Instead, many deaf children begin to learn ASL when they enroll in a school for the deaf or socialize with other deaf children in afterschool programs, camps, or other (in)formal gatherings; others may not learn ASL until well into adolescence or even adulthood (Erting & Kuntze 2008; Morford & Hänel-Faulhaber 2011).

In principle, looking at ASL data on the internet has the potential to spur research that represents the heterogeneous nature of the ASL-signing community more directly. Like other members of the public, researchers have access to any videos of ASL signing that are posted publicly. In practice, however, access to the internet and the drive to post videos to the internet are likely influenced by demographic and socio-economic considerations such as technological literacy and fluency in prestige varieties of ASL. In our estimation, many popular videos are produced by the same group of white, well-educated deaf signers that have been favored by researchers in the past. At the same time, there are other videos showcasing a more diverse pool of deaf signers producing different varieties of ASL, offering researchers a larger sample that is potentially more representative of the ASL community. We expect that diverse representations of ASL signing on the internet, and indeed of sign languages other than ASL, will only increase over time, meaning that in the long term, internet data hold the potential to grant researchers and signing populations unprecedented access to underrepresented language varieties within the ASL-signing community.

A related consideration in analyzing ASL videos from the internet is selecting which videos to work with. It may not be possible to identify a "random" sample of internet data for linguistic analysis. Ultimately, data sampling will depend on the particular research question, weighed against the videos that are available. We propose to minimize "cherry-picking" internet data with the following suggestions, especially for researchers who are interested in issues of frequency, language change, and variation. First, do not consider only an individual signer, but rather a variety of signers, including those who are established vloggers and those who occasionally vlog.[4] Second, consider including all videos from an established vlogger or perhaps a number of videos from a single channel over a certain period of time. In the latter case, consider adopting the standard practice of random data sampling by selecting the videos produced at equal intervals across a time period of the channel, for example, one video from each week for one year. Researchers can also consider contacting the vloggers for demographic background information, to ensure more variety, though this will likely require ethical review approval from the researcher's institution. In the long run, it will also be possible to compare older video sources used in recent publications with newer video sources.

Because ASL videos necessarily reveal the face and body of the signer, researchers who are interested in language variation and change may guess at certain demographic variables of the signer such as age, gender, sex, race, ethnicity, location, and even language background of the signer.[5] However, we urge caution on this front: appearances can be deceiving, and researchers are not immune to language attitudes and ideologies (Hill 2013). It is especially crucial to realize that signers may

"read" as belonging to a particular demographic group, but may identify as members of underrepresented and marginalized minorities including but not limited to African Americans, Native Americans, Latinx/Chicanx, and trans, non-binary, gay, lesbian, and queer persons. Making unfounded assumptions about the background and identity of a signer not only introduces the possibility of inaccurate data, but it also shows a lack of respect for the signer as well as their language and community. Researchers interested in sociolinguistic variation should therefore plan to collect such background information to make their data sampling as principled and ethical as possible.

We also want to comment on genre. Internet-based ASL videos include ASL news sources such as the Daily Moth; vlogs by individual signers touching on a wide range of topics such as politics, car repairs, and cooking; videos sponsored by organizations such as ASLized!; and commercial videos for ASL-signing consumers or for university students who are taking ASL classes. These videos represent a wide range of genres and registers: monologues with and without live audiences, face-to-face and videophone interviews, dyadic and group conversations, rehearsed narratives based on children's stories, and so on. Thanks to inexpensive, accessible video technology, signers are producing a massive number of ASL videos on the internet, resulting in a rich variety of online genres and topics in naturalistic environments. There may even be new genres and subgenres of discourse that have emerged from the interaction of video-recording technological affordances and the visual-manual modality of sign languages. Thus, we encourage researchers who are interested in studying ASL usage on the internet to be mindful of these many genre types.

Finally, researchers should also be aware that vlogs are often subject to some degree of audience design and that vloggers may (metaphorically or literally) edit their content to construct a digital persona. Above all, a primary advantage of internet data is that public ASL videos constitute authentic instances of ASL use, which also encompasses a wide variety of genres. ASL use from the videos should not be considered separate from the use of ASL in "real life"; the internet encompasses part of real life for signers who participate in signing communities online.

3.2 Ethical considerations with internet data

From a researcher's and institutional perspective, videos (and text) that have been posted publicly to the internet, such as vlogs, are generally categorized as "previously collected data" or "no risk studies" and are therefore exempt from institutional ethical review. However, this may vary across educational institutions (Lucas et al. 2013). Given the public nature of internet data, there is, by default, no participant recruitment by researchers. The lack of interaction between the researchers and the video contributors minimizes potential influence from the researcher, including the signers' reaction to the social characteristics of the researchers, on the language data to be studied.

However, researchers need to be aware of potential legal and ethical considerations having to do with video ownership, copyright, and privacy (e.g., Giglietto, Rossi, & Bennato 2012). Videos embedded in *public* posts on social media websites such as Facebook, Instagram, and Twitter do not require an account or log-in for viewing. We recommend that researchers consider only these public posts when working with internet data. However, vloggers may later choose to change their posts to *private*, meaning that they are no longer openly viewable, but rather are shared only with the user's curated list of "friends." Videos embedded in private posts require a user to be friends with the other user who publishes them on their social media page as well as to be part of the same social network to view these videos. Considering these types of private posts as data should involve a comprehensive discussion with the researcher's IRB, including procedures for contacting individuals for permission to analyze their videos.

Relatedly, another ethical concern involves video contributor meta-data. While some contributors disclose information regarding their personal background in their videos, others may only give limited information (e.g., geographical region). As mentioned, given the visual nature of videos, researchers may be tempted to speculate on contributors' age, gender, ethnicity, and other relevant background variables, if contributors do not disclose their identity. We again emphasize the need for caution and reflection on the part of the researcher. If it happens that researchers have insider knowledge on particular contributors, for instance if they are from similar social circles or have mutual connections, then to what extent should the researcher include identifiable information on those contributors? Where do we draw the line? If researchers want to collect demographic information about the signers, they should obtain ethical review approval from

their institution. Although we are experiencing a social transformation of the internet with respect to linguistic and communicative practices in sign languages, there are continuing questions regarding ethical and legal issues of video contributions on different social media platforms, and we as a research community need to continually consider potential changes in ethical practices of collecting internet data.

3.3 Internet data as compared to open data

An additional consideration, related to the discussion of ethical research practices in section 3.2, has to do with copyright ownership as it relates to individual websites' terms of use. There will certainly be legal issues to consider before re-uploading any videos online, including storing videos collected from the internet in any digital data repository (see Collister, chapter 9, this volume). To err on the side of caution, we recommend that researchers contact video creators to request copies of videos and request explicit permission to store the video in an online data repository. This process in turn requires institutional ethics approval as described. We suggest that you consider taking this course of action and have an IRB-approved consent form as soon as possible, because videos may be moved to another website, made private, or taken down at any time, without warning. A permanent link to the videos in a digital data repository will contribute to and promote the practice of reproducible research and open access, making research more transparent and accessible for future scrutiny. This approach is not limited to sign language videos, but any study working with language videos from the internet.

The rise of the internet has also enabled the shift of practices and standards for sharing sign language data: some recently published works incorporate film stills of individual signs along with URLs to the video source from which the signs have been extracted (see Lepic 2019 for a recent example). Other papers are published with selected video clips that are available on a journal-sponsored website (such as *Sign Language and Linguistics*; Zeshan & Panda 2015, for example, provide a link to the video sources used in their article). Still others provide links to the video sources in an online corpus, which may require registration for access; additionally, annotations of the sign language videos may be provided on a digital data repository (see Oomen & Kimmelman 2019 for an example). In each of these cases, the primary data remain accessible to readers for as long as the links remain active. However, there is no guarantee that the links will indeed remain active, and these practices are not yet the norm for most journal publications.

Thus, for the time being, we encourage researchers to provide links to internet-based ASL videos in their conference presentations and scholarly publications, so that other researchers can access and assess the data directly. We also encourage researchers to consider taking steps to deposit their research data, including videos, annotations, and translations, in an open access digital repository. It is not yet the norm in sign language linguistics to make research data, whether elicited in the lab or collected from the internet, available in an archive. However, we hope that this will soon change: the use and citation for internet-based sign language data may help researchers to recognize the benefits of having primary data available for evaluating analyses and for planning future studies. This appreciation may then scaffold data persistence and reproducibility of sign language data sources. This long-term strategy would minimize the ongoing practice of sign language data sets to be short-lived within the limits of a single study.

In the spirit of this push toward using accessible internet data as a bridge to truly open data, we next discuss the handful of small-scale internet-based studies of ASL. Our intention is to highlight the benefits of using internet data for advancing linguistic analyses.

4 Examples of small-scale internet-based studies of ASL

There are a handful of examples of small-scale internet-based studies of ASL. To our best knowledge, Wilkinson was the first investigator to analyze internet-based ASL data, in her corpus analyses on frequency effects on NOT collocations (2016) and the functions of SELF (2006, 2013a, 2013b). The study on NOT collocations included 9.1 hours of internet-based data as a part of the larger data set, which was retrieved during 2006–2007 from the website called DeafRead: Best of Deaf Blogs and Vlogs (http://www.deafread.com). The study investigated the distribution of token and type frequency of NOT collocations, and analysis revealed the three highest-frequency two-sign collocations were identified as [NOT HAVE-TO], [WHY NOT], and [NOT UNDERSTAND]. These phonologically reduced collocations have undergone changes in

semantic-pragmatic function, compared to non-reduced two-sign constructions. This indicated that "signers are not processing sequential relations of two distinct forms, but instead are accessing the chunking unit directly as the collocation has become autonomous in form and meaning" (2016:98), indicating grammaticalization of NOT collocations is taking place in ASL.

The incorporation of internet data in a larger data set also led Wilkinson (2013a, 2013b) to discover genre effects on the usage of three related forms of the sign SELF among American and Canadian signers. The first analysis (2013a) found a robust pattern of SELF usage in vlogs, compared to in narratives and two-person conversations for Americans. The second, variationist study (2013b) compared the American data with Canadian data to explore whether there were differences in patterns of SELF usage among American and Canadian ASL signers. The study identified morphosyntactic variation in ASL in the distribution SELF usage in a variety of genres, revealing that, for example, American vloggers demonstrated a robust preference for employing SELF signs in their vlogs but not in live presentations, whereas Canadian signers showed a more balanced use of SELF forms in vlogs and live presentations.

Two other studies of phonological reduction and morphosyntactic variation in ASL signing on the internet are Lepic (2016) and Lepic (2019). Lepic (2016) documents compound formation processes in contemporary ASL and identifies 104 unique compounds from eighty-seven minutes of ASL signing from fifteen public YouTube channels. These 104 compounds were classified as either *fingerspelled compounds*, which are likely calques of English compounds (e.g., [C-O-N-T-E-N-T QUESTION] "a content question"); *chain compounds*, which are instances of English-ASL bilingual repetition (e.g., [F-I-L-T-E-R FILTER] "filter"); or *sign-sign compounds*, which juxtapose two ASL signs to create a larger unit (e.g., [EXAMPLE SENTENCE] "an example sentence"). Lepic (2016:234) suggests that naturalistic signing on the internet is essential for collecting novel as well as more conventionalized instances of ASL use. This view is also taken up in Lepic (2019), which draws on examples of ASL signing on the internet to examine the gradual erosion of structure in multiword expressions, fingerspelled words, and morphologically complex signs as a function of their frequent use. Many of the ASL examples discussed are linked directly to the relevant video on YouTube, setting the stage for more open and reliable access to (sign) language example sentences in the coming years.

A third internet-based study is Hou, Lepic, and Anible (2018), which investigates the distribution and grammatical functions of the family of LOOK-AT signs in ASL. The data include over eight hundred tokens, from a larger data set of almost fifteen hours' worth of internet-based videos, consisting of eighty-six vlogs from fifty-five unique signers. The data were divided into three broad genre types: conversations (one hour), monologues (four hours and twenty minutes), and broadcast journalism (over nine hours). The sign LOOK-AT corresponds to a prototypically one-handed sign and has been traditionally described as a directional verb that marks the object of visual perception and that codes number and aspect (Klima & Bellugi 1979; Liddell 2003). This sign also participates in other constructions that are literally or metaphorically related to vision; such constructions are often labeled with other English glosses such as ADMIRE, OBSERVE, PERSPECTIVE, and READ, in ASL dictionaries and lexical databases. The English glosses give the impression that these signs are distinct and separate lexical entries. However, looking at the functions within the family of LOOK-AT signs, the investigators find that signers use a variety of LOOK-AT signs in a network of related constructions relating to a wide array of visual and metaphorical perceptions, and these signs exhibit polysemy across all genres. Similar to the studies of Wilkinson and Lepic, the incorporation of internet-based ASL data allowed Hou, Lepic, and Anible to re-examine previous analyses of LOOK-AT signs and to capture the emergence of linguistic structure and meaning among these signs in spontaneous signing across different genre contexts.

What we can learn from the aforementioned studies is that internet-based ASL data have allowed researchers to investigate frequency, language change, and variation and to come up with new analyses that were not previously available using elicited data. Not only have these analyses contributed to the study of sign language linguistics, they have advanced our understanding of sign languages with respect to general linguistic phenomena that have been documented for spoken languages. Finally, these analyses are in principle reproducible, as the data remain open for as long as the links to the videos that were analyzed remain active. The same cannot necessarily be said for the majority of previous ASL studies, in which access to primary data remains quite

closed. The use and re-use of internet data has potential for shifting current research standards and practices toward more reproducible research in sign language research, in light of Berez-Kroeker et al.'s (2018) call for reproducibility in linguistics. Some of the shifts would involve greater transparency about data sources and research methodologies, more direct access to primary data and analyses, and even citations of data sets.

5 Future directions for internet-based (A)SL research

The internet holds a trove of naturalistic data that can be used to chart out new directions in sign language research. An inherent benefit to using internet data is that these videos are authentic examples of ASL as produced by members of the ASL-signing community. While these videos are likely subject to some degree of audience design, there is virtually no risk of researchers compromising the integrity of the available data, as the data are not solicited by experimenters in any way. In addition, videos that have been posted in public spaces online remain visible beyond the time of their original creation and posting. Depending on various factors, many of these videos could remain visible for quite some time.

Internet-based ASL data offer researchers the opportunity to document lexical, grammatical, and sociolinguistic variation from deaf signers across diverse genres and text types, such as monologues and broadcast journalism. Furthermore, internet data also offer fertile opportunities for conducting synchronic and diachronic research. In the case of diachronic research of ASL, one can combine ASL data from the Historical Sign Language Database (Supalla & Clark 2015) and modern ASL signing on the internet to investigate grammaticalization in ASL. Internet data can also reduce and perhaps eliminate the need to conduct metalinguistic elicitation sessions, although this depends on the particular research question. But the potential of internet data is enormous, to the point where researchers can minimize the recruitment and "recycling" of the same language consultants for their studies, while maximizing the opportunity to study underrepresented minority signers within their signing communities.

Internet data also offer researchers the opportunity to document the impact of technology on signing practices, as a growing number of deaf signers are communicating through video chat applications (e.g., FaceTime and Facebook); filming themselves in naturalistic environments such as at their homes, in cars, and in public; and posting their vlogs online (Lucas et al. 2013). Given how the internet has connected signers from near and far in the United States (and their transborder connections also permit them to reach out to other (A)SL signers outside of the United States), the language ecology of the ASL community is changing; signers are broadening their social circles beyond their traditional practices of convening at deaf schools, social clubs, and deaf-oriented events in near proximity to their residences and workplaces. Moreover, as more deaf signers are increasingly encountering each other at international events such as festivals and conferences and through deaf tourism (Friedner & Kusters 2015), signers are also connecting through the internet, making more unprecedented transnational connections possible.

However, there remain a number of challenges related to collecting and managing internet-based ASL videos and in particular sharing and archiving these videos. There are not yet standardized mechanisms for identifying and tagging ASL videos on the internet. We as a field are also still determining the legal and ethical considerations that govern appropriate research practices with internet-based data. In particular, here we have identified the lack of standardized text-based representation systems and the issue of making internet data open and available, beyond the circumstances of individual content creators or website platforms, as tricky problems for which there is not yet a perfect solution.

In this chapter, we have attempted to situate these practical and ethical concerns that stem from studying sign language data on the internet, relative to more traditional methods for obtaining sign language data from language consultants. We are optimistic that as open research and data sharing become increasingly the norm in linguistics, so too will sign language linguists benefit from the push toward open methods.

Notes

1. Indeed, since the very beginning of sign language linguistics, a central concern has been how to best analyze and represent sign language data: William Stokoe's seminal analysis of ASL, for example, demonstrated that individual signs can be described as combinations of structural primes constituting the "tabula" (now typically referred to as *location*), "designator" (*handshape*), or "signation" (*movement*) of the sign, with each structural prime having its own written symbol (Stokoe 1960; Stokoe, Casterline, & Croneberg 1965). However, while

this general approach is still central to sign language analysis, Stokoe's particular symbols are not widely used today.

2. For example, Ted Supalla and others have studied patterns of historical change in ASL, which is made possible by the past efforts of the National Association of the Deaf to preserve videotaped examples of ASL use from 1910 to 1920, as well as the work of deaf filmmakers such as Charles Krauel to document everyday ASL use starting in the 1920s (Supalla 1991; Supalla & Clark 2015).

3. When we say *American Sign Language*, we are referring to signing varieties that emerged naturally among deaf Americans in signing families and residential schools for the deaf and have been transmitted to subsequent generations of deaf people. This is distinct from any artificial signing system such as Signed Exact English invented specifically to teach deaf people the grammar of English. At the same time, we acknowledge the high degree of language contact between ASL and English in the daily lives of deaf people, because signing communities are microcosms of speaking communities, and this presents sign language researchers the challenge of characterizing what constitutes ASL (Lucas & Valli 2010).

4. *Vlog* is short for "video log" or "video blog." A *vlogger* is an internet user who regularly posts vlogs online for others to view.

5. One exception is the Linguistic Video Collection at Gallaudet University, because the videos contain some demographic information about the signers.

References

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56 (1): 1–18. https://doi.org/10.1515/ling-2017-0032.

Crasborn, Onno A. 2015. Transcription and notation methods. In *Research Methods in Sign Language Studies: A Practical Guide*, ed. Eleni Ofanidou, Bencie Woll, and Gary Morgan, 74–88. West Sussex, UK: Wiley-Blackwell.

Erting, Carol J., and Marlon Kuntze. 2008. Language socialization in deaf communities. In *Encyclopedia of Language and Education*, vol. 8, ed. Patricia Duff and Nancy H. Hornberger, 287–300. New York: Springer Press.

Fenlon, Jordan, Adam Schembri, Trevor Johnston, and Kearsy Cormier. 2015. Documentary and corpus approaches to sign language research. In *The Blackwell Guide to Research Methods in Sign Language Studies*, ed. Eleni Orfanidou, Bencie Woll, and Gary Morgan, 156–172. Oxford: Blackwell.

Friedner, Michele, and Annelies Kusters. 2015. *It's a Small World: International Deaf Spaces and Encounters*. Washington, DC: Gallaudet University Press.

Giglietto, Fabio, Luca Rossi, and Davide Bennato. 2012. The Open Laboratory: Limits and possibilities of using Facebook, Twitter, and YouTube as a research data source. *Journal of Technology in Human Services* 30 (3–4): 145–159. https://doi.org/10.1080/15228835.2012.743797.

Hill, Joseph. 2013. Language ideologies, policies, and attitudes toward signed languages. In *The Oxford Handbook of Sociolinguistics*, ed. Robert Bayley, Richard Cameron, and Ceil Lucas, 680–697. Oxford Handbooks in Linguistics. Oxford: Oxford University Press.

Hill, Joseph, and Carolyn McCaskill. 2016. Reflections on the Black ASL Project. *Sign Language Studies* 17 (1): 59–63.

Hou, Lynn, Ryan Lepic, and Benjamin Anible. 2018. When looks count: The function and distribution of LOOK-AT in American Sign Language. Presented at the Sign CAFE 1, Birmingham, UK, July 30.

Klima, Edward, and Ursula Bellugi. 1979. *The Signs of Language*. Cambridge: Harvard University Press.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lepic, Ryan. 2016. The great ASL compound hoax. In *Proceedings of the High Desert Linguistics Society Conference*, vol. 11, ed. A. Healey, R. Napoleão de Souza, P. Pešková, and M. Allen, 227–250. Albuquerque: University of New Mexico.

Lepic, Ryan. 2019. A usage-based alternative to "lexicalization" in sign language linguistics. *Glossa: A Journal of General Linguistics* 4 (1): 23. https://doi.org/10.5334/gjgl.840.

Liddell, Scott K. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge: Cambridge University Press.

Linell, Per. 2011. *Written Language Bias in Linguistics: Its Nature, Origins and Transformations*. New York: Routledge.

Lucas, Ceil, and Robert Bayley. 2005. Variation in ASL: The role of grammatical function. *Sign Language Studies* 6 (1): 38–75.

Lucas, Ceil, Robert Bayley, Ruth Reed, and Alyssa Wulf. 2001. Lexical variation in African American and white American Sign Language. *American Speech* 76 (4): 61–111.

Lucas, Ceil, Robert Bayley, and Clayton Valli. 2001. *Sociolinguistic Variation in American Sign Language*. Washington, DC: Gallaudet University Press.

Lucas, Ceil, Gene Mirus, Jeffrey Levi Palmer, Nicholas James Roessler, and Adam Frost. 2013. The effect of new technologies on sign language research. *Sign Language Studies* 13 (4): 541–564.

Lucas, Ceil, and Clayton Valli. 2010. *Language Contact in the American Deaf Community*. San Diego: Academic Press.

McCaskill, Carolyn, Ceil Lucas, Robert Bayley, and Joseph Hill. 2011. *The Hidden Treasure of Black ASL: Its History and Structure*. Washington, DC: Gallaudet University Press.

Mitchell, Ross E., and Michael A. Karchmer. 2004. Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States. *Sign Language Studies* 4 (2): 138–163.

Mitchell, Ross E., Travas A. Young, Bellamie Bachelda, and Michael A. Karchmer. 2006. How many people use ASL in the United States? Why estimates need updating. *Sign Language Studies* 6 (3): 306–335.

Morford, Jill P., and Barbara Hänel-Faulhaber. 2011. Homesigners as late learners: Connecting the dots from delayed acquisition in childhood to sign language processing in adulthood. *Language and Linguistics Compass* 5 (8): 525–537. https://doi.org/10.1111/j.1749-818X.2011.00296.x.

Oomen, Marloes, and Vadim Kimmelman. 2019. Body-anchored verbs and argument omission in two sign languages. *Glossa: A Journal of General Linguistics* 4 (1): 42. https://doi.org/10.5334/gjgl.741.

Pawley, Andrew, and Frances Hodgetts Syder. 1983. Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics* 7 (5): 551–579. https://doi.org/10.1016/0378-2166(83)90081-4.

Perniss, Pamela. 2015. Collecting and analyzing sign language data: Video requirements and use of annotation software. In *The Blackwell Guide to Research Methods in Sign Language Studies*, ed. Eleni Orfanidou, Bencie Woll, and Gary Morgan, 55–73. Oxford: Wiley-Blackwell.

Stokoe, William C. 1960. *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf*. Buffalo, NY: University of Buffalo.

Stokoe, William C., Dorothy C. Casterline, and Carl G. Croneberg. 1965. *A Dictionary of American Sign Language on Linguistic Principles*. Silver Spring, MD: Linstok Press.

Supalla, Ted. 1991. Deaf folklife film collection project. *Sign Language Studies* 70:73–82. https://doi.org/10.1353/sls.1991.0027.

Supalla, Ted, and Patricia Clark. 2015. *Sign Language Archaeology: Understanding the Historical Roots of American Sign Language*. Washington, DC: Gallaudet University Press.

Wilkinson, Erin. 2006. Does it behave as a reflexive pronoun in American Sign Language? Talk presented at the High Desert Linguistics Society 7 (HDLS 7), Albuquerque, NM, November 9–11.

Wilkinson, Erin. 2013a. A functional description of SELF in American Sign Language. *Sign Language Studies* 13 (4): 462–490. https://doi.org/10.1353/sls.2013.0015.

Wilkinson, Erin. 2013b. Morphosyntactic variation in American Sign Language: A corpus-based investigation on SELF in Canada and the United States. In *Sign Language Research Uses and Practices: Crossing Views on Theoretical and Applied Sign Language Linguistics*, ed. Laurence Meurant, Aurélie Sinte, Mieke van Herreweghe, and Myriam Vermeerbergen, 259–284. Berlin: De Gruyter Mouton.

Wilkinson, Erin. 2016. Finding frequency effects in the usage of NOT collocations in American Sign Language. *Sign Language and Linguistics* 19 (1): 82–123. https://doi.org/10.1075/sll.19.1.03wil.

Zeshan, Ulrike, and Sibaji Panda. 2015. Two languages at hand: Code-switching in bilingual deaf signers. *Sign Language and Linguistics* 18 (1): 90–131.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363