LYNN HOU, RYAN LEPIC, AND
ERIN WILKINSON

Working with ASL Internet Data

Abstract

The internet has become a space where many sign language users post, watch, and share videos and video blogs (vlogs) through social media. This mass of videos produced by, for, and with deaf signers on the internet has yielded a new language ecology that also offers new opportunities for sign language research. We demonstrate how sign language researchers can use the internet for linguistic "fieldwork" by harnessing available, naturalistic sign language data from this online corpus of videos. We discuss a preliminary case study of first-person pronouns based on two vlogs, as a way of offering some practical guidelines for researchers who want to conduct methodologically valid sign language internet-based corpus studies. This case study examines the potential the internet offers for future research advancing the field of sign language linguistics.

The internet holds tremendous potential for scientists interested in language use. Alongside other advances in telecommunication technology in recent decades, the internet has produced a new type of *language ecology* in which users from all over the world interact through a variety of linguistic and communicative practices (Wilson and Peterson 2002). Many internet "spaces," such as digital forums, social media websites, and video chatting platforms, constitute emerging social networks in which new linguistic conventions may arise. This is true not only for globally dominant languages such

Lynn Hou is an assistant professor in the Linguistics Department at the University of California, Santa Barbara. Ryan Lepic is an assistant professor in the Linguistics Department at Gallaudet University. Erin Wilkinson is an associate professor in the Linguistics Department at the University of New Mexico.

as English, but also for underrepresented, minoritized languages like American Sign Language (ASL; Lucas et al. 2013).

Because information posted on the internet typically persists past the moment it is produced and posted, the internet can also serve as a valuable tool for studying the language use that takes place online. This potential has been embraced particularly within the field of corpus linguistics, where the internet is conceptualized both as a direct source of language data ("web as a corpus") as well as a tool for compiling language data ("web for corpus building") (Biber, Conrad, and Reppen 2012; Hundt, Nesselhauf, and Biewer 2013). Again, though not to the same degree as dominant languages such as English, the use of digital technology in language research has also been explored to some extent for sign languages such as ASL (Keating and Mirus 2003; Keating, Edwards, and Mirus 2008; Lucas et al. 2013; Hjulstad 2016).

Considering these two dimensions of the internet, as a new space for language use and as a space for studying new uses of language, we also see the internet as a new space for linguistic fieldwork. We propose that the internet presents language scientists with an unprecedented opportunity to harness naturalistic discourse in ASL for linguistic analysis. Accordingly, the goal of this paper is to offer some practical guidelines of working with internet-based ASL data. We first consider the concept of the internet as a language ecology where naturally occurring, spontaneous practices of signing emerge through video-recording technologies. Then we review the internet as a tool for corpus linguistics of spoken languages, including ethical considerations of working with internet data, and extend the concept of "web as a corpus" to investigating sign languages. Next, we offer some practical guidelines for treating the web as an "ASL corpus," presenting a preliminary study of first-person pronouns in a very small sample of ASL internet data. Finally, we end the paper by proposing future directions for internet-based sign language research in the field of sign language linguistics.

Linguistic Research on the Internet

The Internet as a Language Ecology

The term *language ecology* pertains to the complex web of relationships and interactions that exist between the physical environment,

languages, and their users (Haugen 2001; Mühlhäusler 2003). In referring to the internet as a language ecology, we do not assume that languages are merely used in a particular environment, but rather they have life, purpose, and form; languages exist in dynamic social worlds through the people who use them; and they are characterized by their relationship to society with respect to, social circumstances such as status and intimacy.

The communicative practices of deaf sign language users have been profoundly impacted by the rise of the internet and by an increased availability and affordability of video-recording technology (Valentine and Skelton 2008, 2009; Lucas et al. 2013). What is revolutionary about the internet is that web-based services allow face-to-face communication among deaf people in one-on-one and one-to-many settings over long distances. This contrasts with earlier telecommunication devices known as TDDs (telecommunication devices for the deaf) or TTYs (teletypewriters), which enabled only text-based, one-on-one communication over a telephone line between deaf people. The most obvious constraint of TDDs/TTYs was the lack of face-to-face communication, a requirement for sign language communication: instead, deaf people relied on the written system of the dominant spoken language, which is different from visual signing.

The internet, on the other hand, has enabled the proliferation of videos and vlogs ("video blogs") from ASL signers dispersed across North America and beyond. The availability and affordance of video-recording technology and high-speed wi-fi connections have allowed signers to use and share their sign languages by filming themselves and posting their videos on commercial social media platforms such as Facebook, YouTube, Instagram, and Twitter. These videos encompass a variety of topics ranging from world travels, deaf-related issues, and metalinguistic discourse, in broad genres such as monologues, broadcast journalism, and live presentations. Many vlogs are also produced in response to the mainstream media representations of deaf people and sign languages, or to sociopolitical issues that affect the lives of deaf people. The upshot is that there is a large body of videos in ASL floating around in cyberspace. Thus, the internet can be conceptualized as an emergent language ecology where linguistic and communicative practices in sign languages emerge through existing and potentially new genres and subgenres of discourse.

As an example, on Facebook, there are many social groups, some of which are public, while others can be accessed only by request or invitation. ASL THAT! is a public group "designed for native/fluent ASL users, interpreters, advanced signers (i.e., students who already completed ASL 4 and higher), and CODAs"[1] to discuss and ask questions about ASL signs and ASL interpretations of English words and phrases ("ASL THAT!" 2019). This group has more than 76,000 members and seven administrators; it receives frequent posts with written messages in English and embedded videos in ASL (as of May 2019). One recent post was an inquiry about candidate signs for *Game of Thrones*, a popular fantasy drama TV show with a global viewership. A few responses to the inquiry featured short vlogs to demonstrate different signs signers use or had seen other signers using, and an explanation of how these signs capture the spirit and meaning of the TV show. In this case, we see the potential for language users across a large geographic area to share and comment on one another's language practices in a relatively new digital forum.

Some ASL users are also experienced vloggers who post video content on a regular basis on their own social media profiles. *Deafies in Drag* (https://www.deafiesindrag.com/) is a comedy duo who identify as Deaf, Latinx,[2] and Queer drag queens, and publish their drag performance shows in ASL on YouTube and Facebook. These shows use comedy to call attention to many topics that relate to the experiences of deaf people, such as cultural and language barriers in communicating with the hearing, nonsigning population. The performative practices of hearing drag queens have been analyzed for their patterns of language use and style in English (Barrett 2017), but, to our knowledge, comparable research has not yet been done for ASL. However, we see a new opportunity for studying the communicative practices at work in this new environment, corresponding with the rise of this new ecology for ASL use.

Public vlogs such as those shared on Facebook may also be circulated among larger social networks. MJ Bienvenu, a retired professor in the Department of ASL and Deaf Studies at Gallaudet University, posted a vlog commenting on the purification of ASL on Facebook on January 17, 2018, and subsequently posted another video on YouTube the next day (Bienvenu 2018a, b). Bienvenu's vlog criticized efforts to eliminate initialized signs by replacing them with newer,

invented noninitialized signs. On Facebook, as of May 2019, the vlog received over 256,000 views, 4,000 shares, and 1,100 comments, and on YouTube, the vlog received 3,124 views and eight comments. Bienvenu's vlog also resulted in an interview with the *Daily Moth,* an ASL "radio" show (*Daily Moth* 2018). Although Bienvenu does not appear to be a regular vlogger, this particular vlog nonetheless demonstrates how certain vlogs are widely circulated and shared online, reaching large audiences and amassing thousands of views across space and time.

These few examples demonstrate several types of language use on the internet by members of various ASL-signing communities. This is in addition to other forms of digital communication that are less persistent in nature, such as video messaging applications and programs (Facetime, Marco Polo, WhatsApp, Glide) and video relay services (Convo, Sorenson, ZVRS), which, like telephone conversations, are more ephemeral and private.

Considering these examples, we see the potential for researchers to access populations and topics that have not yet become mainstream in linguistic research on ASL. As Hou, Lepic, and Wilkinson (in press) have discussed, the tradition of linguistic research with ASL-signing consultants in "researcher-controlled" contexts has unfortunately led to a constrained understanding of the structure of ASL due to the very narrow segment of the ASL-signing population included in such studies. By "researcher-controlled contexts," we mean artificial environments such as university laboratories, in which a researcher elicits utterances from a signer, with or without material stimuli, and collects demographic information about the signing consultant. In these contexts, signing consultants are typically "native deaf signers," that is, deaf signers who acquired a natural sign language from birth, who may be well known in their local signing communities, and who may have high levels of formal education and thus, high levels of metalinguistic awareness. However, with the rise of vlogs and videos posted to the internet, we have the potential to study ASL use in "signer-controlled" rather than researcher-controlled contexts, to study phenomena that are not reliably accessible to language users via metalinguistic introspection, and to study the language practices of populations other than the idealized "native deaf signer."

In addition to the opportunity to study sign languages as they are used by signing communities online, the internet also affords more

possibilities for studying naturalistic samples of language usage than have previously been available. Just like the study of spoken languages, such as English, has been dramatically transformed by the development of large, searchable corpora of language data, this same development in the availability of sign language usage on the internet presents an opportunity for ever richer and more naturalistic studies of sign language structure and use. This leads us to discuss how linguists can use the internet as a tool for corpus linguistics, and how sign linguists can likewise use the internet as a tool for sign language research.

The Internet as a Tool for Corpus Linguistics

Corpus linguistics on the web can refer either to accessing a reference corpus online or opportunistically searching the web for linguistic material (Biber, Conrad, and Reppen 2012; Hundt, Nesselhauf, and Biewer 2013). Since the primary interest of our paper is to approach ASL data on the internet as a type of "opportunistic" corpus, the following discussion reviews what constitutes a corpus, starting from the perspective of spoken language data. A corpus is a principled collection of language data: in the Boasian tradition (Jakobson and Boas 1944), in addition to a grammar and dictionary, a corpus of representative texts, including samples of different speakers and genres, is a crucial part of a complete description of a language. Such sets of records are regarded as language documentary corpora (Woodbury 2011).

For some language varieties and projects, a corpus may necessarily be quite limited in size. For example, the Old English Corpus is a complete record of approximately 3,000 surviving Old English texts (https://tapor.library.utoronto.ca/doecorpus/), and the Sources of Early Akkadian Literature corpus contains 880 surviving texts of early Akkadian (https://www.seal.uni-leipzig.de/). These language varieties are no longer spoken, and so a corpus of either variety is a collection of the available texts that remain from when they were used. For contemporary (global) English, on the other hand, several quite large corpora are available, and the number of English texts is continually growing; the News on the Web (NOW) corpus alone automatically adds ~300,000 new web-based newspaper and journal articles every month (https://www.english-corpora.org/now/).

In addition to the size, the contents of a corpus may also vary significantly: corpora may be primarily text-based, in the case of

newspaper article corpora such as the Brown Corpus (http://ota.ox.ac.uk/desc/0668); or they may be made up of transcriptions of recorded spoken conversations, such as the Switchboard Corpus (https://catalog.ldc.upenn.edu/LDC97S62). Beyond their contents, the goals for particular corpora may also differ. For example, a reference corpus such as the British National Corpus (http://www.natcorp.ox.ac.uk/) is intended to be a representative sample of a language variety, in this case British English. In contrast, an opportunistic corpus is more targeted: the EuroParl corpus (https://www.statmt.org/europarl/) contains parallel translations for the proceedings of the European Parliament in eleven languages. The value of these parallel texts as a comparative resource should be weighed against the fact that the text for any single language in this corpus may not be very representative of that language as a whole.

While a "Boasian" corpus as a collection of texts may have once been a physical, printed resource, a modern corpus is also expected to be machine-readable, meaning that users can search the texts in the corpus for any phenomenon of interest. As with the case of the NOW corpus referenced above, most web-based texts are machine-readable by default, which allows researchers to easily mine the web for language data (Lüdeling, Evert, and Baroni 2007).

Another approach is to use a commercial search engine like Google to search for tokens of language use in order to compare frequency estimates of lexical items (e.g., Google Ngrams, https://books.google.com/ngrams) or to identify emerging constructions that may not be attested in existing corpora. Some researchers also treat content from websites such as Wikipedia, Facebook, Twitter, and Reddit as opportunistic corpora. The Indigenous Tweets project (http://indigenoustweets.com/; Scannell 2014), for example, leverages social media to increase the web visibility of minority languages.

We have defined a corpus as a principled sample of language use, and identified both reference corpora and opportunistic corpora as different types of data with slightly different strengths and goals. The internet is an important tool for corpus research into spoken languages, whether for accessing existing machine-readable corpora or for treating web text as a type of opportunistic corpus. However, as researchers, it is also critical that we consider our responsibilities to the digital communities whose language use we wish to study.

Ethical Considerations When Working with Internet Data

In traditional, researcher-controlled contexts, prior to data collection, the researcher must obtain informed consent from the potential language consultant. For researchers affiliated with an academic institution, obtaining informed consent is typically overseen by an Institutional Review Board (see Lucas et al. 2013, 559 for additional discussion about securing informed consent to videotape signers compared to working with available video data). The process of obtaining informed consent involves communicating the goals and methodology of the study and any risk of harm related to participation in the study to the potential consultant, and ensuring that the consultant agrees to participate in light of this information. The reason for this process is simple: researchers have a responsibility to avoid doing harm to others in the course of their research.[3] Researchers working with internet data share the same responsibility to avoid causing harm to others through their research activities (Page, Barton, and Unger 2014). However, there is no "one-size-fits-all" set of regulations for conducting ethical research with internet data (see Markham and Buchanan 2012), and many Institutional Review Boards have not yet updated their procedures to anticipate the types of harm that might arise in working with internet data.

It seems that the biggest risk of harm to others when working with internet data is the risk of *violating an individual's perception of the privacy of the forum in which they are communicating*. Related to this are the potential harms that derive from researchers *making information available about individuals that the individuals themselves did not share publicly*. Researchers therefore must be aware of the perceptions that individual internet users could reasonably be expected to have about the privacy/availability of the content that they have shared online.

To take one extreme example, the data collected for the "Tastes, Ties, and Time" study were publicly released in September 2008. This dataset contained demographic information collected from the Facebook accounts of approximately 1,700 students from a northeastern American university (see Zimmer 2010 for in-depth analysis of this study, and D'Arcy and Young 2012 for further discussion relating to linguistic analysis of Facebook data). The dataset proposed to capture a "snapshot of an entire class over its 4 years in college, including

supplementary information about where students lived on campus, mak[ing] it possible to pose diverse questions about the relationships between social networks, online and offline" (Zimmer 2010, 313). However, as Zimmer (2010) shows, the researchers failed to understand the privacy implications of their project, and the dataset was subsequently taken down. In particular, the researchers combined data they collected from Facebook with housing information and personal email addresses obtained directly from the university. By aggregating and releasing this information, the university and the research team violated the privacy of the students who were the subjects of the study (Zimmer 2010, 322). Moreover, by using research assistants from within the targeted community, the researchers were able to circumvent student restrictions on the accessibility of their Facebook content, and then make that data available to others who were not specifically authorized to access the data, a violation of students' reasonable expectations about the privacy of their Facebook data (Zimmer 2010, 322).

One lesson that we can take from this example is to understand the importance of distinguishing the extent to which studies are analyzing the personal characteristics of *individuals* or analyzing the linguistic properties of *texts*. As researchers, the more we are targeting the demographic variables of individuals, and in particular, collecting personal information about individuals, the more urgently we should feel ethically bound to obtain informed consent from those individuals. Once we consider aggregating information that individuals have not themselves aggregated in a public space, we are ethically bound to allow those individuals authority over the potential dataset. In contrast, the more we are targeting the linguistic features of particular bodies of linguistic data, the less of a privacy risk we pose to the individuals who produced those texts (although this is by no means a clear-cut distinction, as discussed by Page, Barton, and Unger 2014, 60). This means that, especially when working with video data, the researcher should take care not to publicly disclose any additional information about the individual who produced that video that is not already shared as part of that video.

A second lesson from the "Tastes, Ties, and Time" study is to consider the fragile nature of digital privacy, and the expectations that internet users may reasonably have about their privacy online. Many social media websites allow users to restrict who can access their in-

formation through various privacy settings. For example, on Instagram and Twitter, some accounts are publicly available, meaning that they can be viewed by anybody with an internet connection, regardless of whether they have an account on that particular social media platform. However, other accounts are set to "private," meaning that they can be viewed only by individuals who have an account on that social media platform, and who have been specifically authorized by the owner to view the private account. As researchers, the more we target public posts, in particular posts which we believe have intentionally been posted publicly, the less of a risk we pose to the authors of that content. For example, many social media websites allow users to index their content with hashtags, which makes it easier for other users to access and view their content. By analyzing posts that are publicly available, and are indexed with popular hashtags, and furthermore seem to be directed to a wide, public audience, we minimize the risk of violating the expectation of privacy that the author of that post may have of their own content. In contrast, the more we are targeting posts that have been made to "private" groups, or which seem to be in-group conversations that are clearly not intended for a wider audience, the more strongly we are ethically bound to seek informed consent before making the form or content of these posts public (see also Page, Barton, and Unger 2014, 64–72).

We wish to again emphasize that there is no single, uniform set of rules, which can certify that a study is "ethical." Instead, it is our ongoing responsibility as researchers to consider the potential risks of harm that our studies may pose and to minimize those risks as much as possible in collaboration with our institutional ethical bodies and the individuals whose language use we wish to analyze. For our purposes here, by limiting our scope to only those videos that have been posted in public forums that can be accessed without a password, we significantly reduce our risk of violating the expectation of privacy that an internet user may have about the content they have posted online (see also Lucas et al. 2013, 559).

Potential Issues in Using the Internet as a Tool for Studying Sign Languages

Thus far, we have reviewed web-based linguistic research from a primarily spoken/written language perspective. However, in viewing ASL signing on the internet as an opportunistic corpus, we also see a few

points of contrast with existing web-based work. Accordingly, in this section, we expand the idea of "web as a corpus" as it relates specifically to the study of sign languages.

A central issue in working with ASL data on the internet concerns the identification and transcription of videos. Because English is written as well as spoken, a massive amount of English-language data on the internet is already represented in a textual, machine-readable form. Owing to the large market and demand for automatic speech-to-text transcription, even English-language video data can often be automatically (although imperfectly) converted into a written form. For this reason, it is possible to search the internet for constructions of interest, treating the "web as a corpus" for spoken/written languages. However, sign language researchers do not benefit from the use of text in the same way that spoken language researchers do. In short, there are no standardized nor widely adopted writing systems for sign languages (despite numerous attempts to create text representation systems, such as Stokoe Notation, Sutton SignWriting, and the Hamburg Notation System). As a result, researchers cannot simply search the internet for ASL data directly.

Thus, when it comes to sign language videos on the internet, researchers must search for sign language videos, rather than particular linguistic queries, and need to manually annotate the videos they select. Currently, there is no comprehensive library of video sources that represent different sign language varieties. Still, there are a few resources that link to collections of ASL videos for public browsing and viewing. The *Daily Moth* (https://www.dailymoth.com/), for example, uploads all their vlogs on social media (YouTube, Facebook, Instagram, Twitter) and on their own website. Posted vlogs include the recurring segment "Deaf Bing," which features short humorous video clips about the habits and customs of deaf people. Another source, DPAN.TV, "the Sign Language Channel" (https://dpan.tv/), is a video feed that streams ASL videos containing daily news and other accessible information, ranging from music videos to educational programs. They offer a library that includes *DTV News, Tru Biz,* and *The Harold Foxx Show,* to name a few, which registered users can access without cost. DPAN also uses YouTube, Facebook, and Twitter to promote their videos. A third source is ASLized! (http://aslized.org/), an educational

hub that offers video resources on ASL literature, and ASL linguistics, as well as video articles for the *Journal of American Sign Languages & Literatures*. ASLized! also promotes their videos through YouTube and Facebook. Researchers can use such information as a guide for determining the scope and nature of available videos, including whether particular social media platforms allow for users to engage in video-based discussions. Working with a hub like those listed here also allows researchers to easily bypass internet videos posted by novice signers.[4] While the aforementioned websites currently do not offer a search engine for locating videos by category or genre, they typically title their videos in written English and may organize them into collections by subject area. Some videos may be captioned or supplemented with a written transcript, or have a voice-over; while useful, such texts and voice-overs are only an approximate translation at best, and cannot replace the need for fluent signers to manually annotate the linguistic structures in any given video.

The lack of a standardized textual representation for sign languages impacts not only the researcher's ability to find linguistic data, but also the nature of the linguistic analysis. Sign language linguists do share some spoken-language-based practices known as *glossing*, the convention of referring to sign forms with a meaning-based text label (Chen Pichler et al. 2010; Johnston 2010; Crasborn 2015; Hou, Lepic, and Wilkinson, in press). Johnston (2010), for example, identifies "ID-glosses" as unique indices used to identify signs in a citation form ("lemma"). A gloss corresponds to the dictionary meaning of the sign and is represented with a written label from a spoken language like English. The advantage of ID-glosses is that they allow researchers to assign consistent labels to signs in a researcher-controlled corpus and enable researchers to search for previously identified signs.[5] However, ID-glosses require a separate lexical database of individual sign entries in order to facilitate any kind of machine-readable analysis (see Johnston 2010, for more information). As far as we are aware, it is not possible to generate a list of word tokens from a sign language corpus from scratch, as can be done with any text-based spoken language corpus. For all existing sign language corpora, the situation is actually reversed: a database of ID-glosses must be produced *before* the target corpus can even be properly transcribed (Crasborn and Meijer 2012;

Crasborn 2015). In light of these issues, ID-glossing as it is used in researcher-controlled sign language corpus projects may not be completely suitable for web-based linguistic research.

Ultimately, the researcher faces potential problems with whatever sign-representation system they adopt or devise for their analysis. The choice of representation is flexible, in the sense that a researcher can always modify their chosen representation system with additional coding. However, it is also critical to keep the coding representation consistent. We propose that the only sensible solution, for the moment, is to be aware that identical sign forms may nevertheless differ in their linguistic functions in context. We advocate that the analyst consider their sign-representation system as a continuous cycle of reevaluating their coding process according to the needs and developments of their analysis. This recursive and iterative process allows the researcher to make some headway in identifying and representing signs for practical research, without forcing the researcher to determine every aspect of how to represent relevant signs at the outset of the project. We provide an example of how this recursive, iterative coding process might work in the next section.

Working with ASL Internet Data

In this section, we present a "quick and dirty" illustration of how one might go about actually doing linguistic research using ASL data on the internet. We have focused on ASL because the prevalence of ASL data on the internet has already enabled researchers to work with internet-based data in previous studies (Wilkinson 2013, 2016; Lepic 2016, 2019; Hou and Meier 2018; Hou submitted). However, we expect that the methods here will be more broadly applicable to any signing population posting videos to the internet. The following exercise is meant to show some of the benefits and the potential pitfalls of working with internet data and to provide a general workflow that researchers might use to follow or improve upon their own work. More serious work with internet-based ASL data will also require the researcher to adopt different approaches of coding and managing sign language data from the internet, compared to traditional methods of data collection in researcher-controlled settings. Our intention here

is to give interested researchers a methodological template to build from and customize in their own investigations.

Identifying a Question Ahead of Time

Perhaps one of the most promising aspects of studying ASL on the internet is the potential to observe language in use from a variety of signers that is free of the artificial elicitation materials constructed by researchers for use with language consultants, and which also have the potential to directly influence the resulting data. However, it is still necessary to approach ASL data on the internet with specific research questions in mind and to have some expectations about what kind of data can address these research questions. Researchers also must be able to anticipate the types of variation that might be reasonably expected in addressing their question with naturalistic data (the "envelope of variation"). Coding videos is not trivial, and making a flexible plan before starting can save time and effort.

For the purposes of this illustration, we have chosen to examine *the distribution of indexical constructions directed toward the signer*. Because speakers often refer to themselves in discourse, we expect that these constructions will be relatively well attested even in a small sample of data. This makes signer-directed constructions a nice test case for an illustrative methodological sketch. Moreover, as researchers interested in morphosyntactic variation across instances of ASL use, we anticipate that these signs will frequently occur with particular discourse functions, for example, to convey the signer's stance through statements such as "I think" and "I believe." We might also expect that these constructions will exhibit additional unique properties, with differences in communicative function driving differences in their linguistic form. For this reason, we are interested not only in identifying signer-directed constructions, but also in noting the sentential contexts in which they appear.

Accordingly, we initially plan to annotate a window of two signs on either side of each signer-directed construction. We also have reason to expect that some verbs would also participate in signer-directed constructions ("agreement" or "indicating" verbs e.g., Hou and Meier 2018; Fenlon, Schembri, and Cormier 2018); however, for this small

illustration, we will select only independent indexical forms that are directed toward the signer ("first-person pronouns").

Identifying Relevant Data

For this illustration, we want to compare data from two different video sources, as one might when identifying potential differences across genres and/or varieties of ASL. For the sake of a more straightforward comparison, we want to find videos that are approximately the same length and on similar topics. If we do not consider potential sources of variation due to the length of the video or its content, we may overlook factors that could explain any observed differences between data sources, such as keywords that are specific to particular genres or registers. As we will see, this is especially true when working with very small datasets.

We can find public videos by using YouTube and Google to search for particular types of public video content (see figure 1). In order

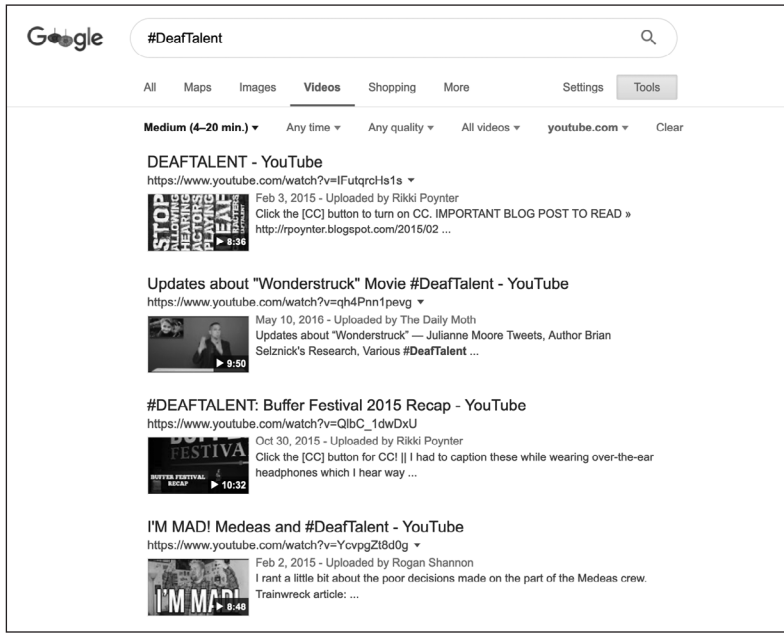

Figure 1. Example of using Google to search for potential ASL videos. Search engines such as Google often have helpful filters for selecting certain types of content. (Image generated on May 1, 2019).

to quickly populate a list of candidate videos, we search for some ASL-specific terms such as "#Deaf Talent," "#WhyISign," "Deaf," and "American Sign Language," each which return several hits. We selected these search terms based on our previous experiences searching for ASL content on the internet; however, different topics and trends will surely be better accessed through different combinations of terms. The hits for our chosen terms include videos that are likely to be good sources for naturalistic ASL data, such as personal vlogs and ASL news videos and videos that are not, such as instructional ASL videos, videos uploaded by novice signers, and English-language videos about ASL.

From these hits, we select two videos, both of which are approximately four minutes in duration, one posted in 2016, and the other posted in 2019 (see table 1). From the video thumbnails, we can attempt to infer some broad demographic information about the signers in each video. The signer in the 2016 video seems to be a middle-aged white woman, and the signer in the 2019 video seems to be an older Black man. It is important to remember that with internet data, social variables such as race and ethnicity, gender, age, socioeconomic class, and the likes are not controlled as much as they are in more traditional data elicited in researcher-controlled settings (see Lucas et al. 2013 for more discussion about the differences between traditional and internet data). It is also important for researchers to remember that the inferences they make about what they see will always be affected by their own implicit attitudes and ideologies (see Lucas and Valli 1992; Hill 2013a, b; Hou, Lepic, and Wilkinson in press; Kusters et al. 2020). For this reason, we keep our inferred variables quite broad, and also explicitly recognize that our inferences are subjectively based on information that is visible to us as viewers. From the titles of the videos, we expect that both signers in the videos we selected will be sharing recollections from their lives. Before we even view these videos, then, we already have a sense of potential similarities (video duration, genre/register) and differences (signer demographics, a general framing of narrative) between them.

Scanning these videos to determine whether we can update or correct our inferences from the title and thumbnail, we also find additional differences between the videos. The 2016 video is more

TABLE 1. Initial Metadata for Videos to Be Analyzed

Upload Date	Thumbnail	Video URL	Video Title	Duration	Inferred Variables*
15-Mar-2016		https://www.youtube.com /watch?v=JPFXWMENf44	#WhyISign	4:34	white, middle-aged, woman, personal, vlog
27-Feb-2019		https://www.youtube.com /watch?v=o13lcyB8qew	Black Deaf History – Joseph Sarpy	4:19	black, older, man, commercial, interview

*After viewing videos, inferred variables may be either modified or elaborated based on the information shared by the video contributor. Importantly, do not solicit the video contributor for further information without first obtaining informed consent after an ethical review.

informal and is shot in a single session, with only signing from the woman pictured in the thumbnail. The 2019 video is a commercial for a video relay service company, so it is heavily edited and is interspersed with several archival photographs. Additionally, we learn from watching the video that the signer in the thumbnail is answering questions from his daughter, who signs in part of the video as well. We determine ahead of time not to analyze this second signer's productions in the video for now.

After this initial inspection of these two videos, we have reason to believe that they will provide an interesting test case for identifying signer-directed pointing constructions in naturalistic ASL. This is because our goal is to illustrate how to begin working with internet data, rather than to provide a definitive analysis of these constructions. After all, eight minutes of video is not much in terms of data. However, we could use the same process outlined above to identify additional videos, or longer videos, for this research question and for additional questions. It is our hope that eventually some aspects of this video selection process will become more automated, but until such a time (and even when the necessary technology becomes widely available), we need to be methodical about selecting videos to study.

Setting up an Initial Coding Sheet

Our coding plan is to identify instances of signer-directed indexical constructions, canonically a point with the index finger to the signer's chest that refers to the signer, in each of the two videos we have selected. Because we are interested in the distribution of these signer-directed constructions, we want to be able to record the two signs on either side of each target construction. This is somewhat similar to "keyword in context" search results, which return concordance lines from a searchable corpus, albeit generated manually. Though annotation tools such as ELAN allow for precise annotation of sign durations (https://tla.mpi.nl/tools/tla-tools/elan/), duration is not one of our main variables of interest here, and we will not mark sign onset and offset in seconds and milliseconds. Although this coding is valuable for many research questions, it can also be a somewhat time-intensive process. Naturally, we feel that researchers should adopt the tools that

will best help them to adequately address their research questions. For now, we want to identify each signer-directed indexical construction in context and to relocate each construction in each video online. For this reason, we set up a simple spreadsheet for recording the target sign and the two signs immediately before or after the target sign, as well as the approximate timecode, in minutes and seconds, which marks where the sign is produced in each video. Table 2 lists the initial coding worksheet columns for this short illustration.

Coders will require a system for identifying signs and sign boundaries, whether based in form, meaning, or a combination of the two. Ideally, these coding systems will be made explicit enough that two separate coders could follow them and arrive at the same coding decisions for the same body of data (the coders and the coding system are then considered "reliable"). Here, we identify individual signs based on the meaning of the utterance in context, which has its own advantages and disadvantages; namely, this process allows for quicker but ultimately more idiosyncratic coding. As we will see next, after an initial round of coding, we also identified formal properties to code for, an instance of the "recursive coding" practice described previously.

Modifying the Coding Scheme after a First Pass

With the initial coding sheet setup outlined above, we are ready for a coding procedure where each line of the database worksheet is an observation of the target construction in our selected videos. An ASL-fluent coder viewed the selected 2016 video at half speed, looking for instances where the signer pointed at or contacted their own chest while referring to themselves (see https://www.signingsavvy.com/sign/ME/3864/1). Once these self-directed constructions were identified, they were labeled as "1p" (mnemonics for this code could be "one finger" or "first-person" or "point") and listed in the fifth column in the coding worksheet, along with the approximate time at which they occurred in the video. Then the coder added information about the signs immediately preceding and following the target construction. In addition to meaning-based glosses, one other idiosyncratic annotation that is used is the double hash (##). Here, this is used when there is a significant perceptible "break," such as when the signer pauses, blinks, stops signing temporarily, and/or recovers

TABLE 2. Column Labels for Initial Coding Worksheet (First Row) with Short Descriptions of the Coding Values in Each Column (Second Row)

Token	onset	tm2	tm1	t	tp1	tp2	VidLink
Sequential count of forms annotated	In MM:SS format, just before the target (approx)	target sign minus 2	target sign minus 1	target sign	target sign plus 1	target sign plus 2	Link to the video on YouTube

Table 3. The First Five Instances of the Target Signer-Directed Construction in the Data

token	onset	tm2	tm1	t	tp1	tp2
1	0:10	##	after	1p	look-at	analyze
2	0:12	look-at	analyze	1p	watch	describe
3	0:13	describe	why	1p	sign5	why
4	0:14	sign5	why	1p	prefer	sign1
5	0:17	why	different	1p	##	make

from a false start. Table 3 lists the first five tokens in context that were identified using this coding procedure.

Already, after only a few seconds of video, we have tricky coding decisions to make. For example, token 2 in table 3 is subtle, at best. It is unclear whether this is actually an instance of the signer indicating and referring to themselves with an identifiable indexical form if the sign glossed as WATCH is being used as an "indicating verb," or if this is an illusion caused by the transitional movement from the sign glossed ANALYZE to the sign glossed WATCH. Similarly, though token 1 and token 5 are very clear, with the signer pointing to the center of their chest with an extended index finger, tokens 3 and 4 are clearly "there," but they are nevertheless very reduced in form: there is no clear index finger extension, and the signer's hand contacts their shoulder, rather than their chest.

In order to keep track of the differences between these tokens, additional form-based coding values are introduced (see table 4). First, overall, is the token "distinct," "reduced," or "unclear," in terms of how closely it resembles the "prototypical" point to the signer's chest with an extended index finger; second, is there any observable English mouthing with the target sign; and last, categorical codes are assigned to the location, handshape, and number of hands used to articulate the target sign. There are many other form-based parameters that could be coded, for example, the movement of the hands or the direction of eye gaze; however, these will not be coded for the time being. Importantly, these coding decisions are, for now, somewhat impressionistic, but bigger questions with bigger datasets will require somewhat more formalized coding criteria in order to be reproducible. While keeping

TABLE 4. Column Labels for Additional Coding Values (First Row) with Short Descriptions of the Coding Values in Each Column (Second Row)

distinct*	mouthing	location	handshape	hands
Code as distinct, reduced, or unclear	Code any identifiable English mouthing, or other	Code either as center or other	Code either as index or other	Code as one, two, or other

*Here, a "distinct" form is defined as a token that closely resembles the prototypical form and is easy to identify. Tokens that exhibit reduction in form are broadly categorized as "reduced." Tokens that are questionable for various reasons are classified as "unclear" and can be left for future investigation.

these future possibilities in mind, this updated coding system can now be used for both of the selected videos.

The only other coding challenge that arises in this small sketch relates to instances in which the signer refers to themselves using a different handshape to accomplish some grammatical function in addition to indicating and referring to the signer. These tokens are coded as target constructions for the same values as outlined above in tables 3 and 4, however, the label they are given is also elaborated to be more descriptive for the communicative function of the sign, such as "1p-my" or "1p-self."

Exploring the Data

Coding the two selected videos following the coding procedure listed above takes little time, only a few hours altogether, including spot-checking and repeated viewings of the videos for tokens that were missed in initial passes. We are now ready to do some initial exploration of the coded worksheet of data.

Altogether, 108 tokens were coded from both videos. Table 5 shows the breakdown of these tokens in a bivariate table, according to the function-based label they were assigned, and whether they were coded as having been articulated in a distinct, reduced, or unclear form. This gives us an initial sense of the data; for example, we can see that 18 coded tokens were "unclear," meaning either that they were so reduced that they were difficult to identify or that they are actually some other type of movement in the signing stream and were

Table 5. Count of Perceived Level of "Distinctness" for Each Signer-Directed Construction in the Combined 8m 53s of Video Data Analyzed

	Distinct	Reduced	Unclear	*Row Totals*
1p	27	43	15	*85*
1p-my	7	8	3	*18*
1p-dual	3	0	0	*3*
1p-formal	1	0	0	*1*
1p-self	1	0	0	*1*
Column Totals	*39*	*51*	*18*	**108**

therefore incorrectly identified as tokens of interest. These unclear tokens can be set aside, and possibly reassessed with a more nuanced coding scheme in the future. This leaves 90 tokens coded as either "distinct" or "reduced." The majority of these 90 tokens are instances of two sign types, the target indicating construction glossed as "1p," and the possessive construction glossed here as "1p-my."

It is important to remember in this coding scheme, that the labels "distinct," "reduced," and "unclear" were assigned somewhat holistically and impressionistically by a single coder. However, we also have more fine-grained coding for phonological variables such as the handshape used to articulate the sign, the location of the sign, and any mouthing with the sign. It would be possible to use these codes jointly to assess potential differences in function and form for the target signer-directed constructions. We could then further refine the coding system as needed. However, due to the relatively small number of tokens being analyzed here, we leave this possibility for the future.

Our 90 "distinct" and "reduced" tokens come from two videos, so we may be curious about any differences between the two videos. When we organize the data of self-directed constructions within each video, as well as the codes assigned to each token, as in table 6, we can see that the majority of tokens (79 percent, 71/90) occur in the video posted in 2016. We can also see that the ratio of distinct to reduced tokens differs across videos: in the video posted in 2016, 69 percent (49/71) of tokens are reduced, and in the video posted in 2019, only 11 percent (2/19) of tokens are reduced.

TABLE 6. Count of Signer-Directed Constructions in Each of the Two Videos that Were Analyzed

	Video Posted in 2016	Video Posted 2019
1p (Distinct, Reduced)	60 (18, 42)	10 (9, 1)
1p-my (Distinct, Reduced)	10 (3, 7)	5 (4, 1)
1p-dual	0	3
1p-formal	0	1
1p-self	1	0
Column Totals	*71*	*19*

What can we make of these potential differences? Unfortunately, the only definitive thing we can say is that it is too soon to make any definitive conclusions. Recall from table 1 that the video posted in 2016 and the video posted in 2019 differ in many ways: the signer in the video from 2016 is a middle-aged white woman signing an informal vlog, and the signer in the video from 2019 is an older Black man participating in an interview that has been edited for a commercial. Even if we were confident that the differences in table 1 are reliable, for example, after statistical analysis such as a chi-squared test or Fischer's exact test, we have no way of determining how much variation might be due to various factors such as differences in age, race/ethnicity, gender, language background, and so on.

However, we also have reason to doubt that the differences between the videos have to do with the signers themselves. Recall that the video from 2019 contains several seconds of archival photographs. Though we selected videos of comparable lengths, the video from 2016 is continuous, unedited signing for the duration of the video, and the video from 2019 has been heavily edited; not only does the video contain elements other than the target signer's signing, but many of the signer's utterances have been "trimmed down," and may not be an accurate representation of how they signed in the moment.

The takeaway for us is that if we wish to follow up on potential differences across varieties of ASL for our chosen construction, we will need to code additional videos that vary with respect to demographics

and other social variables, using this same coding scheme. After all, it is these demographics and social variables—as well as ecologies of language use—that define a community and a language variety. For the time being, we attribute the differences in the overall number and also the proportion of reduced forms to the fact that the video from 2016 is apparently unedited, while the video from 2019 is heavily edited.[6] However, these potential differences in our data will certainly be a useful starting point that can guide our future investigations of this topic.

Investigating the Distribution of Indexical Constructions Directed toward the Signer

After a few hours of work to develop and implement a coding scheme for a small sample of ASL internet data, we have several tokens of our target construction—instances of signing that refer to and are directed toward the signer. Next, we will take only the target indicating ("1p") constructions for further analysis, acknowledging that the majority of these tokens come from only one of our two videos.

Identifying Patterns of "1p" Tokens in the Data

These 70 self-directed "1p" tokens are the construction types we had in mind when developing our original research question: referential points that contact the signer's chest. We expected that these signs will be used in particular sentence contexts, and even in this quite small dataset, this seems to be the case. For example, in our coded data, 1p tokens often occur at the beginning of a signed utterance as the subject of the clause: 16 of the 70 tokens immediately follow a perceptible break in signing ("##," 16 tokens). An additional 13 tokens occur after a "connective" sign signaling the start of a new clause, namely the signs ALSO (5 tokens) and THAT (2 tokens), as well as AFTER, AND, BUT, OK, SO, and UNTIL (1 each), see table 7.

Identifying Functions of "1p" Tokens as Subject Copy

In our database of coded signs, we also find a few examples of the "subject copy" construction, where the self-directed sign appears at the end of a clause and is a co-referential with the previously men-

TABLE 7. Thirteen Self-Directed Constructions in Context; "1p" Tokens that Start a Clause by Virtue of Appearing between a Connective Sign and a Verb or Another Connective. Rows Are Organized Alphabetically according to the Value in the 3rd Column (tm1, "Target Sign Minus 1")

onset	tm2	tm1	t	tp1	tp2
0:10	##	after	1p	look-at	analyze
1:10	and	**also**	1p	say	that
1:21	next	**also**	1p	used-to	1p-my
1:36	1p	**also**	1p	say	1p
1:52	##	**also**	1p	say	1p
2:26	and	**also**	1p	remember	1p
1:36	##	and	1p	also	1p
1:06	not	but	1p	struggle	with
0:59	##	ok	1p	say	that
4:26	##	so	1p	want	share
1:01	say	**that**	1p	used-to	1p
1:11	say	**that**	1p	wish	people
2:29	##	until	1p	go	restaurant

tioned subject of the clause (Padden 1988). An example of a sentence with this construction is seen in example 1, where the second 1p token "re-emphasizes" the first-person subject of the utterance.

EXAMPLE 1. Source: https://youtu.be/JPFXWMENf44?t=21

THAT PA- OLD PAPER,
1p WRITE 3p PRESENT **1p**
"I wrote that old paper for a speech I gave."

To better examine the patterning of the subject copy, we need to extract all instances of this construction from the datasheet and reorganize these tokens by the functions they perform for the verb(s) with which they co-occur. For example, in our database of 70 1p-forms, 7 are self-directed tokens that appear before a coded pause (##) (table 8). In a quick assessment, 2 of 7 self-directed tokens before a coded pause appear to be something other than the subject

Table 8. There Are Seven Instances of 1p Forms Preceding a Coded Pause; Five Instances of Subject Copy Constructions; Two Instances of 1p Forms Expressing Functions Other Than Subject Copy (Highlighted in Gray)

onset	tm2	tm1	t	tp1
0:17	why	different	1p	##
0:24	1p	present	1p	##
0:33	sign5	sign1	1p	##
2:26	1p	remember	1p	##
2:48	school	for	1p	##
3:32	not	understand	1p	##
4:07	that	taught	1p	##

copy construction; one token appears to function as an object of a transitive verb ("taught me"), and another token functions as a dative ("for me"). The remaining 5 of these 7 self-directed tokens appear before a coded pause do seem to be instances of the subject copy construction.

In this dataset, 1p forms functioning as a subject copy co-occur with grammatical categories such as speech act verbs ("present a speech," "signing") and cognitive verbs ("remember," "understand"). However, if we wanted to better understand the subject copy construction in our data, we would need to develop criteria to more directly distinguish subject copies from grammatical objects, across *all* of our coded tokens, in addition to those before a coded pause. For example, we might introduce additional coding for the grammatical function of each 1p token relative to nearby verbs, something along the lines of either "subject," "object," or "other." The next step in exploring this phenomenon would be to examine the distribution of types of constructions that appear in the subject copy utterances. For instance, do we find emerging patterns regarding the use of the subject copy construction with particular types of verbs and predicates? If so, we would then investigate what variables, such as the function of the verb, might lead to differences in patterning for 1p constructions and subject copy constructions. This would again be an instance of flexible and recursive coding practices we have advocated for above, guided by research questions that arise in the course of

analysis, made possible by the relatively broad coding system with which we started.

Identifying Patterns of "1p" Tokens in a Templatic Construction of [1p+verb]

In addition to examples of 1p-tokens that function as either the subject, subject copy, or object of a verb, do we also have larger "1p+verb" constructions that recur with more specialized discourse functions? On the basis of our small dataset, it seems the answer is a tentative yes. Visually scanning the coded datasheet, we find that many 1p+verb collocations involve verbs that either denote speech acts (e.g., "emphasize," "mention," "say," "write") or convey the signer's intents or beliefs (e.g., "prefer," "think," "want," "wish"). Crosslinguistically, these types of verbs are considered likely candidates for taking on additional syntactic functions such as introducing clausal complements, or additional pragmatic functions such as introducing the speaker's view of a proposition in addition to their denoted meaning (Heine and Kuteva 2002; Bybee 2006).

For example, if we compare two instances of 1p followed by the verb LOOK in our coded data, we can see that example 2a is more literal; after the sign LOOK, the signer goes on to describe what they saw on-stage while sitting in the audience. In example 2b, however, the verb LOOK is less literal, as it introduces only the signer's reaction to what they saw, rather than a description of what they saw. In English, this sequence can be rendered with a different construction, such as "and I'm like," which provides an indirect hint to its newer, less literal function in ASL.

EXAMPLE 2A. Source: https://youtu.be/o13lcyB8qew?t=67

AUDIENCE **1p** LOOK,
ALL WHITE, NONE BLACK NONE
"**Watching in the audience, I saw** they were all white, and there was no-one that was Black."

EXAMPLE 2B. Source: https://youtu.be/JPFXWMENf44?t=141

1p LOOK,
fs-AT fs-LEAST fs-HE KNOW SOME SIGN
"**And I'm like**, at least he knows some signs"

This "less-literal" use of the sign LOOK seems to be a straightforward instance of grammaticalization, where a pragmatic inference that is often associated with a construction increasingly becomes part of its conventional meaning. In this case, the inference that we often have reactions to things that we see, or that our description of things that we see also includes our subjective assessment of what we have seen, has actually become part of the meaning of the 1p+LOOK construction in ASL (compare Heine and Kuteva 2002; Bybee 2006; Traugott and Trousdale 2014; and see Hou submitted for a more detailed analysis of this 1p+LOOK(-AT) construction in ASL, in a much larger sample of internet data).

In order to more properly evaluate this potential divergence in function for 1p+verb constructions in ASL, we could introduce another level of coding to our coding system, in the same way that we added columns for additional formal markers in table 4, as well as when we considered additional codes for grammatical function on the basis of examples like those in tables 7 and 8. This process might include adding additional columns, or even color-coding for details of the function (such as whether the verb is a "prototypical" verb or a "nonliteral" discourse marker), and also identifying degrees of formal reduction between adjacent signs (such as a reduction in sign duration or movement). We again leave this as a suggestion for the future, for now (but see Wilkinson 2016 for a more detailed example).

Our final tentative analysis of the 70 "1p" tokens that we have identified is to report which other signs most often follow our target self-directed construction. Consistent with what we might expect from any corpus of natural data, we have a small number of items that occur after "1p" with some frequency (such as LOOK-AT [7 times], GO-FAR [5 times], SAY [4 times], and REMEMBER, WISH, and EMPHASIZE [3 times each]). There is a larger number of verbal signs that follow the self-directed construction "1p" only once in our small dataset (signs such as ENCOURAGE, FALL-IN-LOVE, BE-INVITED, BE-PROUD, and BE-SKILLED).[7] Though we should be cautious not to make too much of these counts on their own, this information might inspire additional future research questions, such as: to what extent do these particular verbs frequently occur with signs other than the self-directed construction we have identified and focused on

here? To what extent do these particular verbs function as prototypes of the larger class of "verbs of saying" and "verbs of desire/intention" that we have impressionistically identified above? Finally, to what extent does the proportion of distinct to reduced forms correlate with constructional frequency of occurrence, in our coded data and in other contexts? These questions can, of course, be addressed through systematic analysis of larger datasets, following the general iterative approach we have outlined here.

Taking Stock

In this "quick and dirty" example of working with ASL data on the internet, we have proposed some strategies for selecting videos and recording metadata about them, as well as steps for developing and iteratively modifying a coding system for answering particular research questions. With the right approach in mind, it is possible to strike a "happy medium" between either focusing in a myopic fashion on only one very limited phenomenon or otherwise being overwhelmed by the sheer number of possible variables and details that might be coded in any sample of naturalistic data. We do not wish to bill this particular system as the perfect system. However, we hope that our contribution can serve as a starting point for other researchers interested in working with signing data on the internet. In particular, we believe that developing a shared sensibility on how to approach a body of data and to develop flexible, recursive coding plans will help to make sense of (and sensible *use* of) the "new frontier" of ASL signing on the internet.

Future Directions for Internet-Based Sign Language Research

Here we have proposed that the internet can be considered a new space for a method of "fieldwork" on sign languages, or perhaps as a "corpus" of naturalistic sign language data. Indeed, ASL usage on the internet causes us to confront many issues inherent to sign language research. Consider the representation of linguistic diversity of ASL signers and their signing varieties in published research to date. With the exception of seminal sociolinguistic studies that found working-class Black signers sign differently than middle-class, white signers (Lucas, Bayley, and Valli 2001; McCaskill et al. 2011), research on ASL structure and use has traditionally targeted only a limited population

of native deaf signers who use a prestigious variety of ASL. This pattern of recruitment follows from particular language ideologies and language attitudes that researchers hold about ASL and its users (Hill 2013a, b). Yet the United States is known for its cultural and linguistic diversity, and this also applies to the deaf signing population, which the field of sign language linguistics has yet to fully embrace for ASL research. In any case, the conservative approach that many 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 an ASL-signing community?* We propose that internet-based sign language videos have the capacity to transform and advance the current scholarship of sign language linguistics for a more comprehensive understanding of language uses and patterns from a larger variety of ASL signers.

It also seems that the internet is functioning as a relatively new type of language ecology, where linguistic and communicative practices in different sign language varieties can emerge. This is in addition to the internet's role as a new area for researchers to describe underrepresented linguistic heterogeneity of ASL signers, and to investigate communities of practice in depth in signing communities. Given the rich diversity of ASL data observed on the internet and the availability of data that has been posted publicly, there is a great potential for future investigations on language usage among ASL signers. We also see no reason why this same perspective cannot also be fruitfully applied to other sign languages. Eventually, it may even be possible, and indeed enlightening, to conduct a comparative study of data from the internet and researcher-controlled contexts, including present-day corpus studies, to better understand potential differences in language use according to the ecologies they occupy. The internet serves as an indispensable source of naturally occurring ASL data, enabling researchers to analyze linguistic structures and communities of practice that have not been previously described in the literature. We encourage other researchers to join us in this new frontier of research.

Notes

1. *Coda* refers to an adult who grew up as a hearing child of (a) deaf adult(s).

2. *Latinx* is an inclusive/nonbinary term for people who identify as a part of Latin American culture and heritage; this term "recognizes the intersectionality of sexuality, language, immigration, ethnicity, culture, and phenotype" (Salinas Jr. and Lozano 2017, p. 9).

3. Fortunately, for many linguistic studies the risks of harm to language consultants are typically quite low, and are usually limited to physical risks such as boredom or fatigue, or social risks such as the risk of compromised privacy after the elicitation session has concluded (see e.g., Hou, Lepic, and Wilkinson in press, Lucas et al. 2013 for additional discussion).

4. Hearing, novice signers are sometimes asked to post videos of themselves signing for their ASL classes, and some novice signers post their attempts to translate popular songs into ASL online. These videos could be analyzed as a type of learner corpus (see Granger, Gilquin, and Meunier 2017), but we recommend avoiding such videos unless the goal of a particular project is to analyze novice signing.

5. An in-development ASL Signbank can be found here: https://aslsignbank.haskins.yale.edu

6. This example illustrates the potential consequences of working with edited videos. They may not be suitable for some research questions, particularly those having to do with the relative timing of linguistic phenomena in a single video. For instance, if one were to investigate the distribution of fingerspelling in a variety of genres, then heavily edited videos might not be appropriate; the act of editing the video might obscure potential gradience in form and function of fingerspelling constructions (cf. Padden and Gunsauls 2003; Lepic 2019).

7. We also attempted to extract the frequency of occurrence for these verbal signs from the literature, namely whether the sign is listed as occurring with a frequency greater than 4.1 per thousand signs in Morford and MacFarlane (2003) or if the sign is listed with a mean subjective frequency rating out of 7 in Mayberry, Hall, and Zvaigzne (2014). Interestingly, these sources differ as to whether they even report frequency information about these particular signs. This underscores the need for additional corpus-based descriptive research in ASL, including Internet-based "corpus" research, and for Internet linguists to pull together different data sources to address the research questions at hand.

References

"ASL THAT!" 2019. Facebook video. *How to Sign Game of Thrones?? ASL??* (blog). 2019. https://www.facebook.com/groups/ASLTHAT/permalink/2437665396464953/.

Barrett, R. 2017. *From Drag Queens to Leathermen: Language, Gender, and Gay Male Subcultures*. New York: Oxford University Press.

Biber, D., S. Conrad, and R. Reppen. 2012. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bienvenu, MJ. 2018a. YouTube video. *Purification of ASL, No!* (blog). 2018. https://www.youtube.com/watch?v=2zXhE5gv3Pw.

———. 2018b. YouTube video. *Response to Vlog Questions* (blog). 2018. https://www.youtube.com/watch?v=k2cSYYfKA24&t=13s.

Bybee, J. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language* 82 (4): 711–33.

Chen Pichler, D., J. A. Hochgesang, D. Lillo-Martin, and R. Müller de Quadros. 2010. Conventions for Sign and Speech Transcription of Child Bimodal Bilingual Corpora in ELAN. *Language, Interaction and Acquisition / Langage, Interaction et Acquisition* 1 (1): 11–40. https://doi.org/10.1075/lia.1.1.03che.

Crasborn, O. A. 2015. Transcription and Notation Methods. In *Research Methods in Sign Language Studies: A Practical Guide*, ed. E. Ofanidou, B. Woll, and G. Morgan, 74–88. West Sussex, UK: Wiley-Blackwell.

———Crasborn, O., and A. de Meijer. 2012. From Corpus to Lexicon: The Creation of ID-Glosses for the Corpus NGT. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, ed. O. Crasborn, E. Efthimou, E. Fontinea, T. Hanke, J. Kristoffersen, and J. Mesch, 13–17. https://doi.org.10.13140/RG.2.1.4007.6884.

"Daily Moth." 2018. Interview with Gallaudet Professor MJ Bienvenu (blog). 2018. https://www.facebook.com/TheDailyMoth/videos/802013746667194/.

D'Arcy, A., and T. M. Young. 2012. Ethics and Social Media: Implications for Sociolinguistics in the Networked Public." *Journal of Sociolinguistics* 16 (4): 532–46. https://doi.org/10.1111/j.1467–9841.2012.00543.x.

Fenlon, J., A. Schembri, and K. Cormier. 2018. Modification of Indicating Verbs in British Sign Language: A Corpus-Based Study. *Language* 94 (1): 84–118. https://doi.org/10.1353/lan.2018.0002.

Granger, S., G. Gilquin, and F. Meunier. 2017. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Haugen, E. 2001. The Ecology of Language. In *The Ecolinguistics Reader: Language, Ecology, and Environment*, ed. A. Fill and P. Mühlhäusler, 57–66. London: Continuum.

Heine, B., and T. Kuteva. 2002. *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Hill, Joseph. 2013a. Language Ideologies, Policies, and Attitudes toward Signed Languages. In *The Oxford Handbook of Sociolinguistics*, ed. R. Bayley, R. Cameron, and C. Lucas, 680–97. Oxford Handbooks in Linguistics. Oxford: Oxford University Press.

———. 2013b. *Language Attitudes in the American Deaf Community*. Washington, DC: Gallaudet University Press.

Hjulstad, J. 2016. Practices of Organizing Built Space in Videoconference-Mediated Interactions. *Research on Language and Social Interaction* 49 (4): 325–41.

Hou, L. Submitted. Looking for multi-sign sequences in American Sign Language.

Hou, L., R. Lepic, and E. Wilkinson. in press. Managing Sign Language Video Data Collected from the Internet. In *Open Handbook of Linguistic Data Management*, ed. A. Berez-Kroeker, B. McDonnell, E. Koller, and L. Collister. Cambridge, MA: MIT Press Open.

Hou, L., and R. P. Meier. 2018. The Morphology of First-Person Object Forms of Directional Verbs in ASL. *Glossa: A Journal of General Linguistics* 3 (1): 114. https://doi.org/10.5334/gjgl.469.

Hundt, M., N. Nesselhauf, and C. Biewer. 2013. *Corpus Linguistics and the Web*. Language and Computers: Studies in Practical Linguistics 59. Amsterdam: Rodopi.

Jakobson, R., and F. Boas. 1944. Franz Boas' Approach to Language. *International Journal of American Linguistics* 10 (4): 188–95.

Johnston, T. 2010. From Archive to Corpus: Transcription and Annotation in the Creation of Signed Language Corpora. *International Journal of Corpus Linguistics* 15 (1): 104–29.

Keating, E., T. Edwards, and G. Mirus. 2008. Cybersign and New Proximities: Impacts of New Communication Technologies on Space and Language. *Journal of Pragmatics* 40 (6): 1067–81.

Keating, E., and G. Mirus. 2003. American Sign Language in Virtual Space: Interactions between Deaf Users of Computer-Mediated Video Communication and the Impact of Technology on Language Practices. *Language in Society* 32 (5): 693–714.

Kusters, A., M. Green, E. Moriarty Harrelson, and K. Snoddon, eds. 2020. *Sign Language Ideologies in Practice*. Sign Languages and Deaf Communities [SLDC] 12. Berlin: De Gruyter Mouton.

Lepic, R. 2016. The Great ASL Compound Hoax. In *Proceedings of the High Desert Linguistics Society Conference*, ed. A. Healey, R. Napoleão de Souza, P. Pešková, and M. Allen, 11:227–50. https://escholarship.org/uc/item/5sf8k4cx.

———. 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.

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

Lucas, C., G. Mirus, J. L. Palmer, N. J. Roessler, and A. Frost. 2013. The Effect of New Technologies on Sign Language Research. *Sign Language Studies* 13 (4): 541–64.

Lucas, C., and C. Valli. 1992. *Language Contact in the American Deaf Community*. San Diego: Academic Press.

Lüdeling, A., S. Evert, and M. Baroni. 2007. Using Web Data for Linguistics Purposes. In *Corpus Linguistics and the Web*, ed. M. Hundt, N. Nesselhauf, and C. Biewer, 7–24. Language and Computers: Studies in Practical Linguistics 59. Amsterdam: Rodopi. https://doi.org/10.1163/9789401203791_003.

Markham, A., and E. Buchanan. 2012. Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0). Unpublished manuscript. http://aoir.org/reports/ethics2.pdf.

Mayberry, R. I., M. L. Hall, and M. Zvaigzne. 2014. Subjective Frequency Ratings for 432 ASL Signs. *Behavior Research Methods* 46 (2): 526–39. https://doi.org/10.3758/s13428–013–0370-x.

McCaskill, C., C. Lucas, R. Bayley, and J. Hill. 2011. *The Hidden Treasure of Black ASL: Its History and Structure*. Washington, DC: Gallaudet University Press.

Morford, J., and J. MacFarlane. 2003. Frequency Characteristics of American Sign Language. *Sign Language Studies* 3 (2): 213–225. http://dx.doi.org/10.1353/sls.2003.0003.

Mühlhäusler, P. 2003. *Language of Environment, Environment of Language: A Course in Ecolinguistics*. London: Battlebridge.

Padden, C. A. 1988. *Interaction of Morphology and Syntax in American Sign Language*. New York: Garland Press.

———., and D. C. Gunsauls. 2003. How the Alphabet Came to Be Used in a Sign Language. *Sign Language Studies* 4 (1): 10–33. https://doi.org/10.1353/sls.2003.0026.

Page, R. E., D. Barton, and J. W. Unger. 2014. *Researching Language and Social Media a Student Guide*. London; New York: Routledge.

Salinas Jr., C., and A. Lozano. 2017. Mapping and Recontextualizing the Evolution of the Term Latinx: An Environmental Scanning in Higher Education. *Journal of Latinos and Education* 18 (4): 302–15. https://doi.org/10.1080/15348431.2017.1390464.

Scannell, K. (2014). Indigenous Tweets. Retrieved from www.indigenoustweets.com.

Traugott, E. C., and G. Trousdale. 2014. Contentful Constructionalization. *Journal of Historical Linguistics* 4 (2): 256–83. https://doi.org/10.1075/jhl.4.2.04tra.

Valentine, G., and T. Skelton. 2008. Changing Spaces: The Role of the Internet in Shaping Deaf Geographies. *Social & Cultural Geography* 9 (5): 469–85. https://doi.org/10.1080/14649360802175691.

———. 2009. 'An Umbilical Cord to the World.' *Information, Communication & Society* 12 (1): 44–65. https://doi.org/10.1080/13691180802158573.

Wilkinson, E. 2013. A Functional Description of Self in American Sign Language. *Sign Language Studies* 13 (4): 462–90. https://doi.org/10.1353/sls.2013.0015.

———. 2016. Finding Frequency Effects in the Usage of NOT Collocations in American Sign Language. *Sign Language & Linguistics* 19 (1): 82–123. https://doi.org/doi 10.1075/sll.19.1.03wil.

Wilson, S. M., and L. C. Peterson. 2002. The Anthropology of Online Communities. *Annual Review of Anthropology* 31 (1): 449–67. https://doi.org/10.1146/annurev.anthro.31.040402.085436.

Woodbury, A. C. 2011. Language Documentation. In *The Cambridge Handbook of Endangered Languages*, ed. P. K. Austin and J. Sallabank, 159–86. Cambridge: Cambridge University Press.

Zimmer, M. 2010. "But the Data Is Already Public": On the Ethics of Research in Facebook. *Ethics and Information Technology* 12 (4): 313–25.