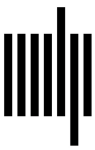
The Open Handbook of Linguistic Data Management

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

21 Managing Sign Language Data from Fieldwork

Nick Palfreyman

1 Introduction: Fieldwork on sign languages in the global South

Language documentation has become increasingly important as a paradigm in linguistic research (Austin 2016), and this is as true for signed languages as for spoken ones. For sign language documentation, however, some of the issues that fall under the heading of data management are rarely discussed and almost never written about (Schembri 2019). Accounts of sign language data invariably jump from collection methods to transcription, annotation, and analysis without stopping for long, if at all, to explain how data are processed, stored, or shared (see, e.g., contributions to Pfau, Steinbach, & Woll 2012; Orfanidou, Woll, & Morgan 2015). The lack of attention to these issues has become ever more conspicuous with the steady growth of literature on sign language documentation and ethics (Fischer 2009; Dikyuva et al. 2012; Kusters 2012, 2015; Nyst 2015; Hou 2017; Hochgesang & Palfreyman forthcoming).

Among a few notable exceptions is the special issue of *Sign Language and Linguistics* (Bergman et al. 2001), which deals with database storage of sign information as well as sign transcription, but naturally some of the details therein have become obsolete over the intervening years. Indeed, most of those documenting sign languages in the field in 2001 were still rewinding and fast-forwarding their way through video cassette tapes or recording signs using notation systems.[1] For the large part, researchers had nothing resembling the multimedia tools or dedicated multitier coding and annotation software that is now in common use. My own earliest experiences of data collection (2010–2011) entailed capturing conversational data onto video cassette tapes—a fact requiring some explanation to the latest generation of researchers familiar only with digital technology.

This case study outlines issues related to processing, storing, sharing, and citing data reported in the literature on sign language documentation. I also draw on my own experience documenting Indonesian Sign Language (BISINDO) from 2010 onward and share a few of the mistakes that I have made along the way. There is some overlap with Crasborn (chapter 39, this volume), but the discussion of data management here reflects some of the particular challenges of being based in the United Kingdom while conducting fieldwork in what is termed the "global South."

2 Sign language documentation and corpora

2.1 Local and distant language documentation

When considering the management of sign language data, it is helpful to make a distinction between two types of documentation that have emerged in sign language research. The first type, which I refer to as *local documentation*, is conducted by researchers based in or near to the community where the sign language is used (Nyst 2015:108), and most of the sign language corpora to have emerged so far fall into this type. Examples include the Auslan (Australian Sign Language) corpus (Johnston 2008), the British Sign Language (BSL) corpus (Schembri et al. 2011), the Corpus NGT (Sign Language of the Netherlands) (Crasborn, Zwitserlood, & Ros 2008), the Corpus Project of Finland's sign languages (Salonen et al. 2016) and the PJM (Polish Sign Language) corpus (Rutkowski et al. 2013).[2]

These corpora usually contain data collected in several regions, but often in relatively controlled environments such as universities and deaf organizations that generate comprehensive records, with good lighting, multiple cameras filming at different angles, and so on (Perniss 2015). Most if not all of these corpora have research teams

that include or are even led by deaf members. The process of liaising with the community is possible because networks are strong, while the community's leaders and at least some of its members are aware of what research is taking place and how it might be important.

The second type, which I refer to as *distant documentation*, is conducted by researchers in countries other than their own and resembles more traditional notions of fieldwork. Linguists of this type have typically been from the global North, documenting sign languages used in non-WEIRD[3] countries of the global South—including Adamorobe Sign Language (Nyst 2012), BISINDO (Palfreyman 2013, 2016, 2019), Inuit Sign Language (Schuit 2013), Kata Kolok (de Vos 2012, 2016), Kenyan Sign Language (Morgan 2017), Malinese Sign Language (Nyst, Magassouba, & Sylla 2011) and San Juan Quiahije Chatino Sign Language (Hou 2016). They usually work with consultants and research assistants, who may need training on linguistics, data collection techniques, ethics, and literacy skills (Nyst 2015).

While the local-distant distinction is not clear-cut, the target language communities for each type are generally rather different. Members of deaf communities in the global South often have fewer resources and less access to education or communication technology for example (Nyst 2015), which might explain why these projects are often led, at least initially, by researchers from the global North. This has important consequences for data management because, as a result of these issues, many researchers have reservations about making data available to other researchers at all (see section 3).

2.2 Types of sign language corpora

A second useful distinction can be discerned in the literature from different emphases that are placed on the attributes of language corpora by researchers from different academic traditions: the first approach is associated with corpus linguists, while the second approach to the corpus is linked with those who identify as documentary linguists. The concerns of corpus linguistics are outlined by McEnery and Wilson (2001:14), who foreground the corpus as representative, finite, machine-readable, and a standard reference. The language documentation approach to the corpus is described by Woodbury (2011:181) using the terms *diverse, ongoing, distributed,* and *opportunistic*, following in the tradition of Franz Boas (Epps, Webster, & Woodbury 2017). Of course, the two approaches are not mutually exclusive, but the distinction is crucial, not least because those creating corpora do not always specify how their corpus relates to these approaches.

Much of the burgeoning work on sign language corpora arguably includes elements from both approaches. On the one hand, considerable effort has been expended on creating machine-readable corpora, with crosslinguistic glossing and annotation conventions emerging to support this (see section 6 for further details). On the other hand, almost all of the sign language corpora developed to date have drawn largely on data produced for the corpus, defying what has been described as a common practice in corpus linguistics of using existing examples of language rather than creating data for the purpose of linguistic analysis (Stubbs 2001:221; Cox 2011:250). One of the few studies that does use existing examples, available online, is described by Hou, Lepic, and Wilkinson (chapter 40, this volume), but paradoxically their corpus is opportunistic and not machine-readable.

2.3 The BISINDO corpus

BISINDO has been used since at least the 1950s (Palfreyman 2019:76), though the language was named in 2006 (from an acronym based on *Bahasa Isyarat Indonesia*, "Indonesian Sign Language") by Gerkatin, the Indonesian Association for the Welfare of the Deaf (Palfreyman 2019:288). The BISINDO corpus comprises nine hours of spontaneous conversational data from 131 participants in six different islands across Indonesia.

Retrospectively, data collection has occurred in two stages.[4] For the first stage (2010–2015), three hours of data were collected from Solo (Central Java) and Makassar (South Sulawesi) using funding obtained from an international non-governmental organization.[5] These data were collected primarily for a comparative study looking at the grammatical domains of completion and negation (Palfreyman 2015), but they were glossed and annotated with the intention of creating a corpus that could be used to answer other research questions.

The second stage (2016–2019) was conducted with funding obtained in 2016 from the Leverhulme Trust.[6] Six hours of data were collected from four locations: Padang (West Sumatra), Pontianak (West Borneo), Singaraja (Bali), and Ambon (Maluku). This was seen as highly desirable to increase geographical representativeness, and the corpus now has a much wider geographic scope.

The choice of field site was also motivated by a desire to reflect better the religious makeup of Indonesia—Pontianak has a high level of ethnic diversity, while the community in Singaraja is mostly Hindu. Singaraja is also the town nearest to the village where Kata Kolok, an unrelated sign language, is used (see section 8), enabling potentially valuable comparisons between a village sign language and the dominant sign language used in the surrounding province. Other considerations include a desire to collect data from places with a relatively long attested history of sign language use (deaf people from Padang were among the earliest from outside of Java to attend the first deaf school, set up in the Dutch East Indies), and the existence of a local community able to work with the researchers.

The BISINDO corpus is best described as an example of distant documentation (see section 2.1), albeit with the intention of moving corpus creation closer to local documentation in future (see section 5). As I mentioned, many sign language corpora encompass elements of both language documentation and corpus linguistics, and the BISINDO corpus is similar in this respect: the corpus is ongoing and opportunistic, with the aim of creating a diverse corpus that is also machine-readable.

3 Informed consent

Informed consent is not always obtained in written format, and researchers such as Austin (2010) have relied on oral consent from hearing people in many endangered indigenous communities. Likewise, for the BISINDO corpus, the conventional method for obtaining informed consent—an information sheet and a consent form—was not used, for three reasons. First, it was difficult to explain to the informants what they were consenting to before they had actually taken part in data collection, because they had no prior experience or understanding of the notion of "research." Second, handing out an information sheet saturated with text is oppressive and inappropriate for a community that has a low literacy rate. Third, an explanation of the research prior to the collection of data would have made it harder to obtain natural data, because this would have prompted unwarranted expectations on the part of informants regarding what was required of them (see Schembri 2008).

With these points in mind, informants who wished to take part were filmed first, and informed consent was obtained afterward, with explanation in sign language as to how the data would be used. This meant that participants now had a clearer understanding as to what they were consenting to. It was hard to explain about the right to withdraw data at a later stage because the informant had only just given consent for their data to be included, so the idea of withdrawal caused some confusion. I therefore informed local deaf community leaders that those who took part in the research could withdraw at any time if they later changed their mind, and that these leaders should let me know if anyone expressed any concerns about their involvement at any stage (Palfreyman 2019:114–115).

Having taken these steps, it is still not possible to say that optimally informed consent has been obtained from all of the participants. In our target communities, some deaf people have not been to school, while those who have been to school had little to no access to the language of instruction. Only a very small number of deaf Indonesians to date have been able to enter further and higher education institutions. While hearing people who have not been to university may be able to draw on general knowledge gleaned from information (over) heard on the radio, the television, or in conversation to understand what might be meant by "research," many deaf Indonesians do not have access to such knowledge.

Although a corpus may be compiled with the aim of linguistic analysis in mind, its participants are sharing their thoughts, stories, experiences, and feelings. Nathan (2011) notes that spontaneous, naturalistic speech "can easily include content that might cause embarrassment, or worse, for the speakers" (112) and cites examples of corpus conversations that reveal illegal activities or damaging statements about other community members. In the BISINDO corpus, there are moments where participants talk about distressing situations, sometimes appearing emotional as they recount what happened, and these can make for uncomfortable viewing. As researchers collecting a large amount of data, we did not observe conversations taking place—to reduce the effects of the observer's paradox—and we did not always have the opportunity to view the data ourselves for several weeks. As a result, we were sometimes not aware of the content of the data before leaving the field.

A longer amount of time in the field may have made it possible to take a different approach: for example, each participant could have been able to review their data and

decide whether they should be permitted. This entails several assumptions, however; for example that each person has the time and inclination to review the data (which is not always the case). It also assumes that participants are able to make decisions based on the range of potential people who might see the data, how they might use them, and how they feel about such people using and viewing them, which is a big ask. As Crasborn (2010) implies, even seasoned researchers cannot know exactly how digital data will be used, especially given ongoing advances in data capacity.

Such caution is not restricted to signed languages (see Gawne & Styles, chapter 2, this volume); spoken language researchers also collect data from informants who may not have a full understanding of what is involved (Thieberger & Musgrave 2007:30–32). The notion of archiving texts that can be accessed via the Internet is not easily understood by people in remote locations with no access to computers. Nathan (2011) notes how those researching endangered languages may regard themselves as having "an ongoing custodial role" (118), controlling access on behalf of their informants.

Sign languages do not have a written tradition (see Crasborn, chapter 39, this volume), and it has been widely noted that video recording is essential when documenting visual-gestural languages (Wilcox 2003; Fischer 2009; Crasborn 2010). The use of space plays a critical role in sign language grammars (Perniss 2012), while facial expressions have important functions at different levels of linguistic organization—including grammar and prosody (Pfau & Quer 2010; Sandler 2012). Full anonymity may therefore seem an impossibility: if documentation efforts are not to lose their value, researchers must retain the faces of their signers, thus increasing the risks associated with making data more widely available (Hochgesang & Palfreyman forthcoming).

4 Responding to the challenge of anonymity

One of the affordances of language documentation is that other researchers can view the primary material on which linguistic analyses are based (Thieberger et al. 2015). There is tension between "formulating, implementing and maintaining access restrictions, and, on the other hand, making materials accessible to the right people for the right purposes" (Nathan 2011:113). In other words, it is important to protect against risks to participants while acknowledging the requirement to build and safeguard academic knowledge.

In response to this challenge, the model that Nathan (2011) proposes, which is based on the Endangered Languages Archive, would be suitable for sign languages. The protocol section allows for several options in terms of who can access corpus content:

- Anyone
- Certain people or groups
  - Research community members
  - Language community members
  - Certain named people or bodies
- Depositor is asked permission for each request
- Only the depositor has access

(117)

These provide a means of "developing further ways for depositors and users to communicate; allowing users to contribute moderated content; and providing detailed reports to depositors detailing accesses of their materials." The aim of such an approach is to encourage a shift in how archives are perceived—as a "dynamic resource at the center of sharing and discussion" (9) rather than simply a set of files. Building up such a community is also a way to ensure that the records can be accessed by users of the language and their descendants (Thieberger et al. 2015).

Johnston (2016) explains how corpus data can be enriched according to the availability of time and resources, by input from successive researchers, who may make annotation passes with similar or different research questions in mind. This approach is used on the BSL corpus website (Schembri et al. 2014) that encourages applications from volunteers and researchers alike.

Another possible solution to the anonymity problem lies in sign language avatar technology, which is becoming ever more sophisticated as research brings together expertise on sign language linguistics, computational linguistics, computer animation, mathematics, and other fields. To date, most of this research focuses on translation between spoken or written texts and a signed language, which is challenging for many reasons (Kipp, Heloir, & Nguyen 2011). Relatively, it is much easier to use this technology simply to replicate text from a human signer on a signing avatar. The ability of avatars to replicate the nuance of sign language production has increased enormously over the past few years, and this creates considerable potential for revisiting the anonymity problem of sign language data.

If software can be programmed to identify manual and non-manual parameters and reconstruct data from signers in avatar form, in a way that captures hand configurations, spatial distinctions, and subtle facial movements such as blinks and eye gaze as well, it may then be possible to treat video data in a way that removes features that identify signers without damaging the precision of representation. Signers and speakers can also be identified by what they say, and content may require additional anonymization. But, as Crasborn (chapter 39, this volume) notes, it is highly likely that automated processing of videos will lead to the identification of phonetic features, and those collecting data should encourage those working on avatars to investigate the application of their work to these ends.

5 Collecting data with the community

There is a considerable literature on conducting research in ways that bring benefits to language communities (Wolfram 1993; Cameron 1998; Benedicto, Modesta, & McLean 2002; Grinevald 2003; Czaykowska-Higgins 2009), and it may be appropriate for researchers to correct erroneous ideas and misconceptions in the community (Labov 1982). Yet, as Nathan (2011:112–113) notes, "by most criteria, the increasing amount of documentation has in itself provided few positive outcomes for communities that want to maintain their languages, or for the evolution of a linguistics discipline that could help them to do so."

Austin (2010:36) draws attention to reciprocity: the researcher should contribute to the community in some way in exchange for the contributions that community members make to the research project. For the documentation of languages used by marginalized or vulnerable communities, researchers might wish to consider how they can use their influence to challenge stigmatized languages (Hochgesang & Palfreyman forthcoming). Certainly, in the case of deaf communities, language documentation and description has been described as "the core activity for sign language vitalization and community empowerment" (Hoyer 2013:43).[7]

Several models have emerged for how linguists can work with deaf communities (see Hochgesang & Palfreyman forthcoming), and Dikyuva et al. (2012) share their experiences as deaf researchers from the global South, offering helpful perspectives on the complex business of negotiating the form that such reciprocity should take. It is all too easy to take an idealistic view of how the relationship with the community should work, perhaps moving from coworking to co-owning, but in practice it is likely that what is possible will be shaped by many constraints. With this in mind, Dikyuva et al. (2012) refer to virtue ethics, which places emphasis on the researcher's moral character, including their ability and willingness "to discern situations with potential ethical ramifications as they arise in the research practice" (Kubanyiova 2008:507). At the very least, one might apply this to data management by paraphrasing Austin (2010:36): "do not manage your data in a way that will make people regret working with you."

Those conducting research in other communities often exchange stories about how they came to enter into the community, and this can be an important consideration from an ethical point of view (Hochgesang & Palfreyman forthcoming). I entered the Indonesian sign community as a volunteer, rather than a researcher, and after working with deaf organizations on capacity building (2007–2009), I switched roles to researcher, documenting BISINDO in a way informed by typology and sociolinguistics (Palfreyman 2019). While I feel fortunate to have worked with the Indonesian sign community for over twelve years, this was not a planned course of action, at least from the start. In my case, the research journey has been shaped by factors as varied as who I volunteered with, funding applications (some accepted, others rejected), and chance encounters, both with certain academics and with informants who wanted to become more involved in research.

Having gained experience of documentation in the first stage of data collection (see section 2.3), data for the second stage were collected in partnership with Pusat Penelitian Tuli (PUPET), a social research foundation set up in 2014 with Muhammad Isnaini, a deaf Indonesian man. Isnaini had assisted with transcription in the first stage and went on to work with me on data collection and transcription for the second stage. This brought knowledge and experience of documentation in-country, placing Isnaini in a position to pass this on to other interested persons, especially deaf members of the sign community wishing to document their own language (further details about this approach appear in Hochgesang & Palfreyman forthcoming).

6 Choice of metalanguage and other data processing decisions

The glossing and enrichment of sign language data entails many decisions, not least concerning the use of annotations to specify different types of signs—such as lexical signs, fingerspelling productions, pointing signs, gestures, depicting signs, and so on (chapters in Pfau, Steinbach, & Woll 2012 give a helpful overview to these and other types of signs). It makes good sense to follow common conventions, developing, challenging, and adapting as necessary, and the most comprehensive annotation guidelines for sign language data to date were created for the machine-readable Auslan corpus (Johnston 2016). A recent project, Digging into Signs, builds on this by identifying annotation standards that are emerging crosslinguistically (Crasborn, Bank & Cormier 2015), while the Global Signbank (signbank.science.ru.nl) offers a standardized template that those documenting a sign language can use to create a database of lexical entries.[8]

These guidelines are specifically crafted for the purpose of creating a machine-readable corpus and lexical database, and this kind of work continues to be both labor-intensive and time-consuming. The BISINDO corpus team currently has three members (including the author) working on it part time, which is far from optimal. That said, technological fixes continue to appear that aim to automatize processes such as creating video clips and organizing signs according to sublexical parameters, which promises to accelerate the business of organizing and annotating sign language corpora in due course.

Another issue requiring more attention is the lack of guidelines for documentary corpora of sign languages (section 2.2). For situations where researchers lack the resources to make a corpus machine-readable, or are not aiming to compile a machine-readable corpus, guidelines would be valuable, and to continue without such guidelines as the field of sign language documentation expands will most likely result in ever more fragmented sign language data management practices.

A further issue concerns the selection of a metalanguage for glossing. Instead of allocating a random alphanumeric code for each sign, it is common to use glosses or labels: these are invariably words from a metalanguage that make it possible to analyze sign language corpora (Johnston et al. 2011:12). For sign language documentation, the metalanguage is almost always a written language with which the sign language has contact (Lucas 2013).

For the BISINDO corpus, the metalanguage for glosses and free translation is Indonesian, which "enhances the accessibility for the research consultants and for future researchers in Indonesia—most of whom will not have a good command of English, but will have a working knowledge of Indonesian" (Palfreyman 2019:99). With find-and-replace functions and annotation programs such as ELAN (Sloetjes 2014) now offering multiple language options, it will then be possible to translate the metalanguage into English with relative ease, which enables greater access for the international academic community.

Indeed, the Global Signbank manual (Crasborn et al. 2018) specifies that a parallel gloss must be created in English, to enable cross-corpora analysis. Once again, however, a balance must be struck between ensuring that the corpus can be accessed by the international academic community and by the community that uses the target language, which is challenging for several reasons. Academics often face more pressure, from funders for example, to prepare data for other academics than for the sign or speech communities concerned. Further to this, when working with sign communities with low literacy levels, there are issues around the accessibility of the metalanguage. One of the ways to make the metalanguage more accessible is to set up links between glosses and a lexical database (such as the Lexicon service linking ELAN with the Signbank) so that written glosses can be viewed as signs, but this does not solve the accessibility of annotations and the like.

7 Learning about sign language data management

Many of those documenting sign languages at postgraduate level and beyond will have learned more about linguistics and comparatively much less about data management, although research culture is changing swiftly in this area. Bad practices that take root in one's early research tend to persist if unchallenged, and for many good reasons, it makes sense to remain attuned to developments in data management. While a few research teams are dedicated to sign language documentation, it is still common for sign language researchers to find themselves in university departments that focus on spoken language research, and the ensuing isolation is compounded in the

case of deaf researchers who do not always enjoy access to information.

Summer schools can be an excellent way to learn, although there is usually a heavy bias toward spoken languages, with course leaders who do not always remember that some languages are signed rather than spoken. Very occasionally, sessions deal specifically with sign languages: for example, on several recent occasions at the Leiden Summer School in Linguistics.

Online forums are an excellent place to seek advice from those with experience of sign language documentation, and it is worth posting queries to groups such as Deaf Linguists, who are usually very happy to share their own practices. Conferences also offer an opportunity to ask for advice: presentations dealing with technical aspects of data management are still lamentably rare, but it can only be a good thing for researchers to open up and share their approaches to managing data. In my experience, issues around sign language documentation have arisen informally during conversation at international gatherings such as the SIGN conference series and the Sign CAFÉ workshops.

Optimal or recommended technical specifications for compatibility with annotation software such as ELAN should be available in the user manuals, including file type (.mp4, .wmv, and so on), frame height, and frame rate, and these can be useful to ensure that recordings are compatible with the software. It is sensible to look for advice on preferred recording formats and settings elsewhere in the documentary linguistic literature, and/or in training institutes and online forums, so that recordings are in line with current best practice recommendations for long-term media preservation and reuse. The technical requirements of language archives must also be met. In sum, planning ahead is always beneficial: data recorded in the wrong format will have to be converted, and converting large numbers of files unnecessarily is best avoided where possible, especially when working in the field.

The consequences of files with incompatible specifications may not always be immediately apparent, but I have had experiences where the same video files have worked with annotation software on one computer, while producing indecipherable output on another computer. Unfortunately, on at least one such occasion the latter output was in Indonesia and I was in the United Kingdom, which held up data analysis considerably. Despite a plethora of data-sharing options, it remains difficult to share very large files internationally, and sound planning—making sure that the right files are left securely with the right team member in the right country—really does pay dividends.

8 Data management for sign language documentation: Future directions

Perhaps the most inevitable future direction for language documentation entails collecting and managing data from the Internet, now that more signers are posting videos on Facebook, Instagram, and other social media platforms in different sign languages (including BISINDO). As the Internet creates communicative spaces and transforms real-life practices, it generates immense opportunities as well as dilemmas that researchers need to engage with, linked for example to ethics and the collection of metadata (see Hou, Lepic, & Wilkinson, chapter 40, this volume).

The use of data from sign communities that are in contact with "deaf tourism," and other language contact situations, is another promising area for sign language documentation. This is especially pertinent for Indonesia because Bali receives a regular stream of deaf tourists from around the world, and several deaf-led enterprises have emerged to cater for them (Moriarty Harrelson 2019). Two major projects examining contact situations, Sign Multilingualism (Zeshan & Webster 2019) and Deaf Communication without a Shared Language[9] answer theoretical questions using data mostly collected in laboratory settings, but thus far few data have been collected in situations where languages are naturally in contact.

As described in section 2.1, filming in the laboratory has obvious and important benefits—including control over lighting and the use of numerous cameras—but also limitations, especially for those wishing to elucidate the use of language in situ. The Kata Kolok corpus (de Vos 2016) is arguably richer in this respect than most others, as all data were collected in and around the village where the language is used. While the BISINDO corpus also comprises data filmed in situ, Kata Kolok signers were also recorded in a range of cultural contexts, such as informal gatherings and religious ceremonies (de Vos 2016:211). I conclude by suggesting that we need more of this kind of in situ language documentation: these data offer multiple insights to linguists, sociolinguists, and anthropologists and offer an important

counterbalance to data collected in the controlled settings of the laboratory.

Notes

1. As with many spoken languages, sign languages have no widespread written form and nothing resembling the International Phonetic Alphabet (Nyst 2015). Rudimentary systems such as Stokoe Notation and SignWriting have been developed to encode the sublexical components of signs, but they are not in common use in most countries, and the availability of video recording has replaced the use of such notation.

2. A publicly available corpus for American Sign Language (ASL) has not yet been created; several documentation projects are underway—including the ASL Signbank (Hochgesang, Crasborn, & Lillo-Martin 2019), which provides a collection of ASL signs linked with identification glosses—but these are not sign language corpora.

3. WEIRD stands for Western, educated, industrialized, rich, democratic.

4. Due to the funding required to collect data from sites across such a vast country, the expansion of the corpus was highly desirable but far from certain during the first stage. The general dearth of long-term funding highlighted by this case is, unfortunately, quite common, and made it harder to plan the corpus in advance.

5. CBM (Christian Blind Mission), an organization for disability-inclusive development.

6. Leverhulme Trust Early Career Research Fellowship, ECF-2016–795.

7. Many of the world's deaf people continue to face discrimination; lack of access to sign language, education, and information, in particular, create disparities and make deaf people more vulnerable than their hearing counterparts (for more information see Hochgesang & Palfreyman forthcoming).

8. One of the main aims of the Signbank is to facilitate cross-linguistic comparison.

9. This project is run by Prof. Onno Crasborn at Radboud University 2017–2022.

References

Austin, Peter K. 2010. Communities, ethics and rights in language documentation. In *Language Documentation and Description* 7, ed. Peter K. Austin, 34–54. London: SOAS.

Austin, Peter K. 2016. Language documentation 20 years on. In *Endangered Languages and Languages in Danger: Issues of Documentation, Policy, and Language Rights*, ed. Luna Filipović and Martin Pütz, 147–170. IMPACT: Studies in Language and Society 42 Amsterdam: John Benjamins. https://doi.org/10.1075/impact.42.02gri.

Benedicto, Elena, Dolores Modesta, and Melba McLean. 2002. Fieldwork as a participatory research activity: The Mayangna linguistic teams. In *Proceedings of the Twenty-Eighth Annual Meeting of the Berkeley Linguistic Society* 28 (1): 375–386.

https://journals.linguisticsociety.org/proceedings/index.php/BLS/article/view/3852.

Bergman, Brita, Penny Boyes-Braem, Thomas Hanke, and Elena Pizzuto, eds. 2001. Sign

transcription and database storage of sign information. Special issue, *Sign Language and Linguistics* 4 (1/2).

Cameron, Deborah. 1998. Problems of empowerment in linguistic research. *Cahiers de I'ILSL* 10:23–38.

Cox, Christopher. 2011. Corpus linguistics and language documentation: Challenges for collaboration. *Language and Computers* 73:239–264.

Crasborn, Onno. 2010. What does "informed consent" mean in the internet age? Publishing sign language corpora as open content. *Sign Language Studies* 10 (2): 276–290.

Crasborn, Onno, Inge Zwitserlood, and Johan Ros. 2008. *The Corpus NGT: A Digital Open Access Corpus of Movies and Annotations of Sign Language of the Netherlands*. Nijmegen, the Netherlands: Centre for Language Studies, Radboud Universiteit Nijmegen.

Supplementary material: http://hdl.handle.net/hdl:1839/00-0000-0000-0004-DF8E-6. ISLRN: 175-346-174-413-3.

Crasborn, Onno, Inge Zwitserlood, Els van der Kooij, and Anique Schüller. 2018. *Global Signbank Manual*. Version 1. Nijmegen, the Netherlands: Radboud University, Centre for Language Studies.

Crasborn, Onno, Richard Bank, and Kearsy Cormier. 2015. Digging into Signs: Towards a gloss annotation standard for sign language corpora. Technical report. http://doi.org/10.13140/RG.2.1.2468.5840.

Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous communities. *Language Documentation and Conservation* 3 (1): 15–50.

de Vos, Connie. 2012. Sign-spatiality in Kata Kolok: How a village sign language of Bali inscribes its signing space. PhD dissertation, Max Planck Institute of Psycholinguistics, Nijmegen.

de Vos, Connie. 2016. Sampling shared sign languages. *Sign Language Studies* 16 (2): 204–226.

Dikyuva, Hasan, Cesar Ernesto Escobedo Delgado, Sibaji Panda, and Ulrike Zeshan. 2012. Working with village sign language communities: Deaf fieldwork researchers in professional dialogue. In *Sign Languages in Village Communities*, ed. Ulrike Zeshan and Connie de Vos, 313–344. Berlin: de Gruyter & Ishara Press. https://doi.org/10.1515/9781614511496.313.

Epps, Patience L., Anthony K. Webster, and Anthony C. Woodbury. 2017. A holistic humanities of speaking: Franz Boas and the continuing centrality of texts. *International Journal of American Linguistics* 83 (1): 41–78.

Fischer, Susan. 2009. Sign language field methods: Approaches, techniques, and concerns. In *Taiwan Sign Language and Beyond*, ed. James H.-Y. Tai and Jane Tsay, 1–19. Chia-Yi, China: Taiwan Institute for the Humanities, National Chung Cheng University.

Grinevald, Colette. 2003. Speakers and documentation of endangered languages. In *Language Documentation and Description*, vol. 1, ed. Peter K. Austin, 52–72. London: SOAS.

Hochgesang, Julie A., and Nick Palfreyman. Forthcoming. Sign language corpora and the ethics of working with the community. In *Sign Language Corpora*, ed. Jordan Fenlon and Julie A. Hochgesang. Washington, DC: Gallaudet University Press.

Hochgesang, Julie A., Onno Crasborn, and Diane Lillo-Martin. 2019. *ASL Signbank*. New Haven, CT: Haskins Lab, Yale University. https://aslsignbank.haskins.yale.edu/.

Hou, Lynn Y-S. 2016. "Making hands": Family sign languages in the San Juan Quiahije community. PhD dissertation, University of Texas at Austin.

Hou, Lynn, Y-S. 2017. Negotiating language practices and language ideologies in fieldwork: A reflexive meta-documentation. In *Innovations in Deaf Studies: The Role of Deaf Scholars*, ed. Annelies Kusters, Maartje De Meulder, and Dai O'Brien, 339–359. Oxford: Oxford University Press.

Hoyer, Karin. 2013. *Language Vitalization through Language Documentation and Description in the Kosovar Sign Language Community*. Nijmegen, the Netherlands: Ishara Press. www.oapen.org/download?type=document&docid=442947.

Johnston, Trevor. 2008. *Auslan Corpus*. London: SOAS, Endangered Languages Archive. https://elar.soas.ac.uk/Collection/MPI55247. Accessed April 12, 2019.

Johnston, Trevor. 2016. *Auslan Corpus Annotation Guidelines*.

https://media.auslan.org.au/attachments/Johnston_Auslan-CorpusAnnotationGuidelines_February2016.pdf.

Johnston, Trevor, Adam Schembri, Kearsy Cormier, Jordan Fenlon, and Ramas Rentelis. 2011.

Type/token matching in annotated SL corpora: Examples from Auslan and BSL corpus projects. Presentation at the workshop Building Sign Language Corpora in North America, Gallaudet University, Washington, DC, May 21–22.

Kipp, Michael, Alexis Heloir, and Quan Nguyen. 2011. Sign language avatars: Animation and comprehensibility. In *Intelligent Virtual Agents. IVA 2011*, ed. Hannes H. Vilhjálmsson, Stephan Kopp, Stacy Marsella, and K. R. Thórisson, 113–126. Lecture Notes in Computer Science 6895. Berlin: Springer. https://doi.org/10.1007/978-3-642-23974-8_13.

Kubanyiova, Maggie. 2008. Rethinking research ethics in contemporary applied linguistics: The tension between macro- and microethical perspectives in situated research. *Modern Language Journal* 92 (4): 503–518. https://doi.org/10.1111/j.1540-4781.2008.00784.x.

Kusters, Annelies. 2012. Being a deaf white anthropologist in Adamorobe: Some ethical and methodological issues. In *Sign Languages in Village Communities: Anthropological and Linguistic Insights,* ed. Ulrike Zeshan & Connie de Vos, 27–52. Berlin: de Gruyter.

Kusters, Annelies. 2015. *Deaf Space in Adamorobe: An Ethnographic Study in a Village in Ghana*. Washington, DC: Gallaudet University Press.

Labov, William. 1982. Building on empirical foundations. In *Perspectives on Historical Linguistics*, ed. Winfred P. Lehmann and Yakov Malkiel, 17–92. Amsterdam: John Benjamins.

Lucas, Ceil. 2013. Methodological issues in studying sign language variation. 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, 258–308. Berlin: de Gruyter & Ishara Press. https://doi.org/10.1515/9781614511472.285.

McEnery, Tony, and Andrew Wilson. 2001. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Morgan, Hope. 2017. The phonology of Kenyan Sign Language (Southwestern Dialect). PhD dissertation, University of California, San Diego. https://escholarship.org/uc/item/9bp3h8t4.

Moriarty Harrelson, Erin. 2019. An ethnography of deaf tourist mobilities. Presented at ASL Lecture Series, University of Pennsylvania School of Arts and Sciences, November 20.

Nathan, David. 2011. Archives 2.0 for endangered languages: From disk space to MySpace. *International Journal of Humanities and Arts Computing* 4 (1-2): 111–124. https://doi.org/10.3366/ijhac.2011.0011.

Nyst, Victoria. 2012. *A Reference Corpus of Adamorobe Sign Language: A Digital, Annotated Video Corpus of the Sign Language used in the Village of Adamorobe, Ghana*. Leiden, the Netherlands: Leiden University Centre for Linguistics.

Nyst, Victoria. 2015. Sign language fieldwork. In *Research Methods in Sign Language Studies: A Practical Guide*, ed. Eleni Orfanidou, Bencie Woll, and Gary Morgan, 107–122. Malden, MA: Wiley Blackwell.

Nyst, Victoria, Moustapha Magassouba, and Kara Sylla. 2011. *A Digital Annotated Video Corpus of the Local Sign Language used in Bamako and Mopti, Mali*. Leiden, the Netherlands: Leiden University Centre for Linguistics.

Orfanidou, Eleni, Bencie Woll, and Gary Morgan, eds. 2015. *Research Methods in Sign Language Studies: A Practical Guide*. Oxford: Wiley Blackwell.

Palfreyman, Nick. 2013. Form, function and the grammaticalization of completive markers in the sign language varieties of Solo and Makassar. In *Tense, Aspect, Modality and Evidentiality in Languages of Indonesia*, ed. John Bowden, 153–172. NUSA 55. http://hdl.handle.net/10108/74331.

Palfreyman, Nick. 2015. Sign language varieties of Indonesia: A linguistic and sociolinguistic perspective. PhD dissertation, University of Central Lancashire.

Palfreyman, Nick. 2016. Colour terms in two Indonesian sign language varieties: A preliminary analysis. In *Semantic Fields in Sign Languages*, ed. Ulrike Zeshan and Keiko Sagara, 269–300. Berlin: de Gruyter and Ishara Press. https://doi.org/10.1515/9781501503429-008.

Palfreyman, Nick. 2019. *Variation in Indonesian Sign Language: A Typological and Sociolinguistic Analysis*. Berlin: de Gruyter Mouton. http://doi.org/10.1515/9781501504822.

Perniss, Pamela. 2012. Use of sign space. In *Sign Language: An International Handbook*, ed. Roland Pfau, Marcus Steinbach, and Bencie Woll, 412–431. Berlin: de Gruyter.

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

Pfau, Roland, and Josep Quer. 2010. Non-manuals: Their grammatical and prosodic roles. In *Sign Languages*, ed. Diane Brentari, 381–402. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511712203.018.

Pfau, Roland, Markus Steinbach, and Bencie Woll, eds. 2012. *Sign Language: An International Handbook*. Berlin: de Gruyter. https://doi.org/10.1515/9783110261325.

Rutkowski, Pawel, Joanna Lacheta, Piotr Mostowski, Joanna Filipczak, and Sylwia Lozinska. 2013. The corpus of Polish Sign Language (PJM): Methodology, procedures and impact. Presentation at Research, Records and Responsibility: Ten Years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures, December 2–4. https://ses.library.usyd.edu.au/handle/2123/13310. Accessed December 19, 2019.

Salonen, Juhana, Ritva Takkinen, Anna Puupponen, Henri Nieminen, and Outi Pippuri. 2016. Creating corpora of Finland's sign languages. In *Workshop Proceedings of the Seventh Workshop on the Representation and Processing of Sign Languages: Corpus Mining / Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, ed. Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, and Johanna Mesch, 179–184. Paris: European Language Resources Association (ELRA).

Sandler Wendy. 2012. Visual prosody. In *Sign Language: An International Handbook*, ed. Roland Pfau, Marcus Steinbach, and Bencie Woll, 55–76. Berlin: de Gruyter. https://doi.org/10.1515/9783110261325.55.

Schembri, Adam. 2008. The British Sign Language Corpus Project: Open access archives and the observer's paradox. Presentation at Workshop on Construction and Exploitation of Sign Language Corpora, LREC, Marrakech, Morocco, May 26–June 1.

Schembri, Adam. 2019. Making visual languages visible: Data and methods transparency in sign language linguistics. Presentation at TISLR13, Hamburg, September 27.

Schembri, Adam, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier. 2011. *British Sign Language Corpus Project: A Corpus of Digital Video Data of British Sign Language 2008–2011*. 1st ed. London: University College London. http://www.bslcorpusproject.org.

Schembri, Adam, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier. 2014. *British Sign Language Corpus Project: A Corpus of Digital Video Data and Annotations of British Sign Language 2008–2014*. 2nd ed. London: University College London.

Schuit, Joke. 2013. Signs of the Artic: Typological aspects of Inuit Sign Language. PhD dissertation, University of Amsterdam.

Sloetjes, Han. 2014. ELAN: Multimedia annotation application. In *The Oxford Handbook of Corpus Phonology*, ed. Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 305–320. Oxford: Oxford University Press.

Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Thieberger, Nick, and Simon Musgrave. 2007. Documentary linguistics and ethical issues. In *Language Documentation and Description*, vol. 4, ed. Peter K. Austin, 26–37. London: SOAS.

Thieberger, Nick, Anna Margetts, Stephen Morey, and Simon Musgrave. 2015. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36 (1): 1–21. https://doi.org/10.1080/07268602.2016.1109428.

Wilcox, Sherman. 2003. The multimedia dictionary of American Sign Language: Learning lessons about language, technology and business. *Sign Language Studies* 3 (4): 379–392.

Wolfram, Walt. 1993. Identifying and interpreting variables. In *American Dialect Research*, ed. Dennis Preston, 193–221. Amsterdam: Benjamins.

Woodbury, Anthony C. 2011. Language documentation. In *The Cambridge Handbook of Endangered Languages*, ed. Peter K. Austin and Julia Sallabank, 159–186. Cambridge: Cambridge University Press.

Zeshan, Ulrike, and Jenny Webster, eds. 2019. *Sign Multilingualism*. Lancaster, UK: Ishara Press and de Gruyter.

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