39 Managing Data in Sign Language Corpora

Onno Crasborn

1 Introduction

With the arrival of video on desktop computers at the start of this century, the use of data in the study of signed languages saw substantial changes. Where before, few larger data sets were created (primarily for observational data in first-language acquisition), from 2004 onward many research groups started constructing sign language corpora (Crasborn et al. 2007; Johnston 2009). These corpora were virtually all elicited, rather than harvested from external sources, as such sources simply did not exist in sufficient quantities. Even today, it is debatable whether YouTube and similar online platforms contain a large enough collection of the use of a particular sign language from which a balanced sample could be created (Crasborn & van Winsum 2014). In addition, the available metadata are often limited and legal and ethical issues also make it difficult to actually use such data for research purposes. An exception is the TV-recorded weather reports that were interpreted from spoken German to German Sign Language (Deutsche Gebärdensprache, or DGS), and collected in a corpus by Rheinisch-Westfälische Technische Hochschule Aachen for the purpose of developing sign language technologies (Bungeroth et al. 2006). Such a limited domain of language use has its advantages for developing automatic sign recognition and machine translation of sign to speech. Because the topic of language use is constrained, the "type-token ratio" of signs will be better (fewer signs occurring only once, for instance).

The corpora that are based on elicitation generally serve a wide variety of purposes in linguistic research and also for applied uses. They form the first large-scale documentation efforts for the sign languages involved, enable research in a variety of linguistic domains, allow for the development of corpus-based lexicography, and are used by teachers and learners. Because of the lack of a written tradition for any sign language, film or video is crucial for each of these domains and consequently, sign language corpora are having major impact on research and are promising to have a long-term impact on the language communities as well (see Gawne & Styles, chapter 2, this volume).

This impact comes almost directly from the documentation of the language: for the first time, both laymen and experts have access to a record of the language in all its present-day diversity. Lacking a commonly used writing system and the concomitant lack of a written tradition, a sign language "library" in the past two decades consisted of printed dictionaries with photos and a set of CD-ROMs and DVDs, if at all. In the construction of sign language corpora, age is always a parameter (e.g., see Johnston & Schembri 2006; Crasborn & Zwitserlood 2008; Schembri et al. 2013; Bono et al. 2014): we know from observation and some research (e.g., Frishberg 1975 and Supalla & Clark 2014 on American Sign Language) that sign languages throughout the world have changed over the last century and continue to change in part due to educational and language policies. In that sense, present-day sign language corpora are "snapshots" of the language: they are recorded at one point in time, rather than harvested from sources that may span multiple decades. There is not at present a monitor corpus of any sign language that keeps recording new data as time progresses.

More indirectly, the availability of annotated corpora allows for the development of language technologies such as machine translation that are becoming ever more important. Currently, many deaf and hard of hearing users are using spoken language technologies such as automatic speech recognition to aid in specific situations. The development of sign language recognition

and synthesis has certainly made progress (see, for example, the collection of presentations from the most recent Conference on Language Resources and Evaluation, or LREC, workshop on the representation and processing of sign languages; Efthimiou et al. 2020), but is far from being ready for a wide range of applications (Crasborn 2010a). The need for manual transcription and annotation is hampering the development of the very large data sets that are used by state-of-the-art machine learning techniques.

In terms of research data management, the relatively young history of multimodal annotation and archiving and the privacy-sensitive nature of video make sign language corpora different from text and audio corpora and similar to audiovisual data sets collected to study nonverbal behavior. This chapter aims to discuss some of the data management issues that the developers of sign language corpora are faced with. The different parts of the data life cycle, from data collection to sharing are discussed in turn, focusing on the use of ELAN for transcription and annotation (used for most sign language corpora).

2 Data collection

Video data are the primary data for all sign language corpora, sometimes supplemented by infrared (Kinect) or three-dimensional video recordings to obtain depth information (Jayaprakash & Hanke 2014). While for more experimental studies, motion tracking is also used (and can sometimes be displayed and analyzed by the same software as where the annotations are created), this has not been done for larger sign corpora yet.

This inevitable use of video as primary data calls for extra attention in the protection of privacy (see Kung, chapter 8, this volume). Anonymization of the video images themselves is impossible without destroying the value of the video recordings (Crasborn 2010b). Given the importance of facial expressions in signed interactions (Baker-Shenk 1983; Crasborn 2006; Herrmann & Steinbach 2011), it is important that people's faces are not only recognizable but also recorded at such an orientation and resolution that fine facial movements can even be detected when signers rotate their head in various directions. This implies a separate camera on each of the interlocutors and either a high-resolution camera (like the present-day high definition, 1920×1080 pixels) or a dedicated camera zoomed in on the head. Either way, good lighting is imperative to record high-quality video. In terms of research data management, two approaches can be taken. Research participants can be asked to give informed consent for public distribution, acknowledging that this means eternal availability for anyone and permitting for uncertainty as to what future technologies will be able to do with the data (Crasborn 2010b). Although public, "copyleft" licenses are getting more and more common, it is debatable whether researchers will actually be able to enforce the conditions of the license (such as the Creative Commons conditions "no commercial use" and "share alike"); see also Collister (chapter 9, this volume) for further discussion. Alternatively, the video data are shielded off from public use and shared only with researchers who sign a restrictive end user license. This limits the potential impact on and use by any given sign language community, but makes it more likely that signers can agree to their recordings being archived and reused.

As with audio(visual) corpora of spoken languages, some of the metadata that make the data most interesting for research use concern person variables such as age, gender, and linguistic skills (Trippel 2004). For signers, given the heterogeneity of the deaf community in terms of language background, further information about family history can be crucial: Did their parents and grandparents use sign language? Were they deaf or hearing? For archives (including those such as the Language Archive [TLA] of the Max Planck Institute for Psycholinguistics that house sign language collections), it is important that metadata are all public so that data sets can be located. However, the combination of all the signer properties that researchers may wish to have access to are so specific that they will easily lead to unique individuals in the (often small) language communities, even without having the video. Following discussions in a workshop on the documentation of sign language heritage ("Metadata for Sign Language Corpora" workshop of the European Cultural Heritage Online [ECHO] project, Nijmegen, the Netherlands, May 8–9, 2003), an elaborate and a reduced set of metadata categories was considered, the latter being recommended as a standard (Crasborn & Hanke 2003). This proposed standard included information about the deafness of participants and their parents, for instance, as well as the age of acquisition of their sign language. Further details about their hearing status, such

as the use of hearing aids or the precise hearing loss in decibels, were considered unnecessary and too privacy-sensitive, especially because metadata are often publicly accessible even if the data are not. More detailed information can still be included in an archive when protected under a (more restricted) end user license. If the metadata profile permits, corpus-level statistics can also be published independent of the individual videos, stating, for instance, that 85% of the signers in the data set acquired the primary sign language under the age of six years. What is considered sensitive information may differ between communities, and the involvement of deaf signers (whether as informants, assistants, or researchers) is of key importance for large as for small language communities (cf. Pollard 1992; Harris, Holmes, & Mertens 2009; Singleton, Martin, & Morgan 2015).

Many sign language corpora collected in the last fifteen years are remarkably similar, which importantly promotes reproducible research across corpora and languages (Gawne & Styles, chapter 2, this volume). They tend to focus on dialogues rather than monologues and multilogues, where narratives are also recorded in a dialogue setting. Narrative tasks such as recounting fable stories or the Tweety and Sylvester cartoon "Canary Row" complement more interactive tasks such as discussions of deaf-related themes and relatively free interaction. While this makes for comparable data sets across sign languages, it also restricts our view of language use in deaf communities (cf. Good, chapter 3, this volume). The narrative tasks are not necessarily the most ecologically valid samples of language use. Increasingly, we see collections of more spontaneous data resulting from fieldwork being archived in the same way as the sign language corpora of Australian Sign Language (Auslan), Sign Language of the Netherlands (Nederlandse Gebarentaal, NGT), British Sign Language, and other signed languages.

These interactions are typically recorded by multiple video cameras, combining frontal views of signers with a view of the whole scene and top shots or zoomed-in recordings of the face. Recordings tend to use dark plain backgrounds and have signers wearing fairly plain clothing without too many patterns, contrasting well with their skin color. One specific technical point of attention is use of manual focus, as automatic focusing in many cameras can be ineffective with moving hands at short range. The resulting video data form the primary data for a sign language corpus, and the quality of these recordings is therefore paramount. While for some annotation and analysis tasks low-resolution video may suffice, for others the full frame of standard definition or high definition will be important. For instance, if there is only one camera recording a frontal view of each signer, fine facial movements may not be visible in a reduced frame. Details of body positions and movements are better visible at higher temporal resolutions and with the use of good lighting and professional cameras: the in-camera compression used by cheaper cameras is especially detrimental in poor light conditions. The current rise of 4K video (quad high definition) at high frame rates (fifty or sixty frames/second) promises a great step forward for future sign corpora, with a concomitant impact on storage space. Although it may still be a challenge for annotation software to play back multiple synchronic video streams at such high resolutions, it is no longer necessary to create low-resolution working copies of primary data in addition to full-resolution archive copies (see also Mattern, chapter 5, this volume).

3 Data processing

Multimodal annotation of sign language data is not fundamentally different from that of other types of multimodal data; there is more convergence in the tools that are used within the sign language linguistics community. The large majority of researchers use the open-source stand-alone tool ELAN developed by TLA at the Max Planck Institute for Psycholinguistics (Wittenburg et al. 2006),[1] with a few exceptions of research groups that use the proprietary server-based software iLex, developed by the University of Hamburg to integrate the annotation of discourse with a lexical database (Hanke, Rodriguez, & Paz Suarez Araujo 2002).[2]

A general requirement of annotation workflows applies just as much to sign language data sets as to any other type of annotation: to use clear and systematic setups of annotation documents (using different tier types for different types of information), separating annotations for individuals on different tiers, using participant labels in ELAN, and so forth. Figure 39.1 illustrates how this has been approached in the Corpus NGT.

ELAN cannot enforce the application of a template across a corpus, so extra attention is needed and possibly some scripting to find inconsistencies and remove these.

Figure 39.1
An annotation file for the Corpus NGT, illustrating how tier names use "S1" and "S2" to refer to the signer on the left and the right, respectively, in a dialogue. The pop-up menu shows further tier information, including a participant code that uniquely identifies each individual.

Explicit annotation guidelines and the use (or development) of a lexical database are vital especially for glossing, as there is no shared sign language orthography that is in common use by any deaf community nor linguistics community. The use of "ID glossing" has been extensively discussed in the literature surrounding sign language corpora (Johnston 2008), and the annotation guidelines of the Auslan corpus (Johnston 2016) have served as a model for all other sign language corpora.[3] A Digging into Data project in 2015 led to a summary of shared glossing practices between the British Sign Language and NGT corpora, which highlights the categories sign language researchers are likely to want to distinguish in the manual channel (Crasborn, Bank, & Cormier 2015). Figure 39.2 shows how different gloss tiers have been set up for the left and right hands in the Corpus NGT, enabling independent alignment of annotations for the activity of the left and right hands (see Crasborn & Sáfár 2016 for further discussion of this annotation scheme).

The transcription and annotation of non-manual channels is generally not done across the board, with the exception of mouth actions (Crasborn & Bank 2014). The intensive code mixing that can be observed in the use of mouthings (mouth actions stemming from spoken words) brings many lexical elements into interactions that will inform linguistic analyses at other levels. The transcription and annotation of features such as eye gaze, eye brow states, or nose wrinkles appear to have a much worse trade-off between effort and general benefit. These are then added to specific segments when a corpus is actually used for dedicated linguistic studies.

As ELAN saves its information in XML documents (ELAN Annotation Format; extension .eaf) and ELAN is still being actively developed and maintained, basic future compatibility is ensured. However, the version of the ELAN Annotation Format file format has been changing slowly over the years, and old annotation files sometimes need to be updated to work with (all features of) new versions of the software. At present (version 6.1), ELAN does not offer an automated way to update files for a whole corpus of dozens or even thousands of files. For this and other corpus-wide processing, the development of scripts that work on large batches of annotation documents can be useful. The integrated and server-based approach offered by iLex clearly has its advantages in this respect.

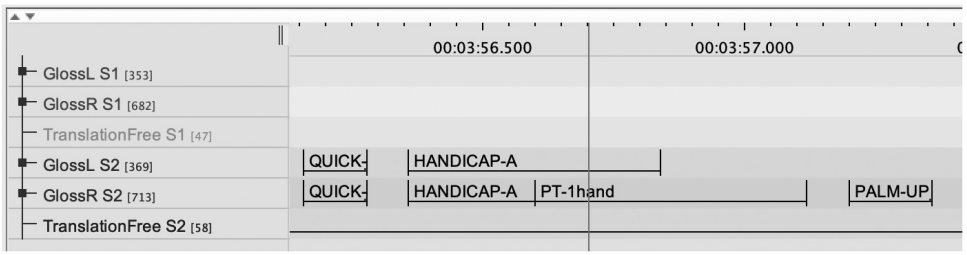


Figure 39.2
Independent alignment of glosses for each hand enable the transcription of so-called spreading behavior, where one hand of a two-handed sign (here, HANDICAP-A) is held while the other hand articulates the next sign.

A larger concern in using archived data lies in any lexicon links that have been used. The online lexicon will likely have evolved after archiving the annotation data, and inconsistencies of various types might arise. More generally, the use of external controlled vocabularies (containing list of values for a certain annotation layer) that are not part of the archive but keep changing if they are used for new data sets and adapted because of new insights can lead to such inconsistencies. Values may have disappeared from the vocabulary or acquired a new meaning. It is therefore recommended to also archive vocabularies along with the annotation files. For a Signbank lexicon or older LEXUS lexicons (the two types that ELAN 6.1 can link to), it is at present not clear how an archived version could be created. As a minimal option, a comma-separated values dump of the lexicon could be archived along with the annotation files. Although not a user-friendly option for researchers who want to use the archived data, archiving text versions of external controlled vocabularies (which are already in XML format) at least ensures interpretability of the data.

For the Corpus NGT, there are now four public releases of the annotation files, which complement the archiving of the video files in TLA in 2008.[4] These successive releases not only include a step-by-step growth of the number of annotated files, but also corrections to obvious errors, and changes to glosses that came about during relemmatization of the Global Signbank data set for NGT since 2008.

4 Storing

There are two primary archives that currently host sign language data sets: the Endangered Languages Archive presently includes the Auslan corpus and data from twelve other sign languages, and TLA hosts the Corpus NGT and data from some ten other sign languages. Some other sign language corpora are stored on local university servers. The two large archives share metadata standards: earlier ISLE Metadata Initiative (IMDI), and now Component Metadata Initiative (CMDI) files are required to archive data. The European Common Language Resources and Technology Infrastructure (CLARIN) project and similar language resource projects have contributed to this converging standard. The extension to the IMDI standard can be flexibly implemented in a CMDI profile, with room for variation.

Global Signbank, an outgrowth of the NGT Signbank that was based on the Auslan Signbank software (Cassidy et al. 2018), currently hosts data sets from eleven different sign languages, most of them in their initial stages of language documentation, corpus creation, or historical research across sign languages. It also hosts shadow copies of ASL Signbank (English; American Sign Language), VGT Signbank (Dutch; Vlaamse Gebarentaal), and an English/French LSFB (Langue des Signes de Belgique Francophone) lexical data set to facilitate research on "international sign," the highly dynamic language contact practices between users of different sign languages.

Although the software of the different Signbank systems all go back to the original one for Auslan, the details of the systems vary. In particular, the degree to which morphological information is encoded and the level of detail in the phonological description differs. This can make direct comparison of data sets rather difficult. In this sense, and in the lack in registration in larger databases of language resources, the Signbank data sets are still far from the FAIR (Findable, Accessible, Interoperable, and Reusable) principles.

5 Sharing

Although anonymity is difficult if not impossible to maintain in video-recorded interactions of signed languages, many researchers are opting for open access publication following the choice of the Corpus NGT to be open access in 2008. They feel that the difficult if not endangered position of signed languages in many countries calls for maximal visibility. The limited availability of video resources for language learning and the lack of written resources make for a potentially large impact of open access corpora. The Corpus NGT, for instance, is widely used in deaf education, for training sign language interpreters and sign language teachers, and for entry-level courses to NGT. The first sign language corpus created for Auslan was made with funding support of the Endangered Languages Documentation Programme following a successful case for the endangerment of Auslan (Johnston 2004). Sign language corpora can thus be seen as documenting languages under pressure, but they can also be seen as instruments for language (re)vitalization (McKee & Manning 2015). Ethical concerns around the publication of a person's data (see Holton, Leonard, & Pulsifer, chapter 4, this volume) are taken seriously but, with people's explicit consent, are mitigated by the need for data inside and outside the academic world.

6 Future perspectives

The biggest difference between corpora for signed as opposed to spoken language corpora lies in the need for manual annotation. There is currently no equivalent of automatic speech recognition that could aid in the basic transcription of sign language use. This is likely to change rapidly in the coming decade. Researchers recording and archiving new sign language corpora might therefore want to prioritize the collection of more primary data over the annotation of those data: once automated processing of videos will lead to, first, phonetic features and later, with advances in machine learning, tokenization of manual signs, data sets can be processed and made available for linguistic research that are much larger than the thirty to three hundred hours of video that we see nowadays. Interestingly, the increase of (semi)automated annotation may well alleviate our present concerns with privacy of signers in video recordings, as many studies will become possible based on transcriptions of signed interactions without the need for access to the original videos.

7 Conclusion

Sign language corpora have had an enormous impact on sign language linguistics. With the lack of a writing system and the late arrival of technology to record, store, and share recordings of signers, only now have linguists been enabled to do research on the basis of published data sets. At the same time, many aspects of the technologies involved are still under development. Although the ELAN annotation tool has become a de facto standard, it lacks many of the advanced corpus management features that the proprietary tool iLex has. The creation of lexical databases dedicated to sign language data is currently seeing rapid development, and here too, only since very recently have such lexical data sets become available for use by a wider research community. In the coming decade, further developments are expected that will impact research data management, including the improved integration between ELAN and the Signbanks, the FAIR publication of lexical data sets, and the addition of new data coming from automated analysis of videos using computer vision and pattern recognition.

Resources

For an overview of sign language corpora, see the survey of the DGS-Korpus team at Hamburg University: https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/sl-corpora.html.

Notes

1. https://tla.mpi.nl/tools/tla-tools/elan/.

2. https://www.sign-lang.uni-hamburg.de/ilex/.

3. More information can be found at http://www.auslan.org.au/about/annotations/. The latest version of the annotation guidelines is published on https://mq.academia.edu/TrevorJohnston.

4. https://hdl.handle.net/1839/00-0000-0000-0004-DF8E-6.

References

Baker-Shenk, Charlotte L. 1983. A microanalysis of the nonmanual components of questions in American Sign Language. PhD dissertation, University of California, Berkeley.

Bono, Mayumi, Kouhei Kikuchi, Paul Cibulka, and Yotaka Osugi. 2014. A colloquial corpus of Japanese Sign Language: Linguistic resources for observing sign language conversations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, ed. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, H. Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 1898–1904. Paris: ELRA.

Bungeroth, Jan, Daniel Stein, Philippe Dreuw, Morteza Zahedi, and Hermann Ney. 2006. A German Sign Language corpus of the domain weather report. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, ed. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2000–2003. Paris: ELRA.

Cassidy, Steve, Onno Crasborn, Henri Nieminen, Wessel Stoop, Micha Hulsbosch, Susan Even, Erwin Komen, and Trevor Johnston. 2018. Signbank: Software to Support Web Based Dictionaries of Sign Language. In *Proceedings of LREC 2018*, ed. Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, Takenobu Tokunaga, 2359–2364. Paris: ELRA.

Crasborn, Onno. 2006. Nonmanual structures in sign languages. In *Encyclopedia of Language and Linguistics*, 2nd ed., ed. Keith Brown, vol. 8, 668–672. Oxford: Elsevier.

Crasborn, Onno. 2010a. The Sign Linguistics Corpora Network: Towards standards for signed language resources. In *Proceedings of the 8th Conference on Language Resources and Evaluation (LREC)*, ed. Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, 457–460. Paris: ELRA.

Crasborn, Onno. 2010b. What does "informed consent" mean in the Internet age? Publishing sign language corpora as open content. *Sign Language Studies* 10 (1): 276–290.

Crasborn, Onno, and Richard Bank. 2014. An annotation scheme for the linguistic study of mouth actions in sign languages. In *Beyond the Manual Channel: 6th Workshop on the Representation and Processing of Sign Languages*, ed. Onno Crasborn, Eleni Efthimiou, Stavroulou-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, and Johanna Mesch, 23–28. Paris: ELRA.

Crasborn, Onno, Richard Bank, and Kearsy Cormier. 2015. Digging into Signs: Towards a gloss annotation standard for sign language corpora. Project deliverable, Nijmegen, the Netherlands, and London. https://www.ru.nl/sign-lang/projects/completed-projects/digging-signs/.

Crasborn, Onno, and Thomas Hanke. 2003. Additions to the IMDI metadata set for sign language corpora. Unpublished manuscript, Radboud University. http://sign-lang.ruhosting.nl/echo/docs/SignMetadata_Oct2003.pdf.

Crasborn, Onno, Johanna Mesch, Dafydd Waters, Annika Nonhebel, Els van der Kooij, Bencie Woll, and Brita Bergman. 2007. Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics* 12 (4): 535–562. doi:10.1075/ijcl.12.4.06cra.

Crasborn, Onno, and Anna Sáfár. 2016. An annotation scheme to investigate the form and function of hand dominance in the Corpus NGT. In *A Matter of Complexity: Subordination in Sign Languages*, ed. M. Steinbach, R. Pfau, and A. Herrmann, 231–251. Berlin: Mouton de Gruyter.

Crasborn, Onno, and Frouke van Winsum. 2014. NGT online: A first inventory. Poster presented at Exploring New Ways of Harvesting and Generating Sign Language Resources: Legal, Technical, and Crowd-Sourcing Issues, CLARIN workshop, Hamburg, December 13–14, 2014. https://www.ru.nl/sign-lang/events/past-events/clarin-workshop/.

Crasborn, Onno, and Inge Zwitserlood. 2008. The Corpus NGT: an online corpus for professionals and laymen. In *Construction and Exploitation of Sign Language Corpora: 3rd Workshop on the Representation and Processing of Sign Languages*, ed. Onno Crasborn, Eleni Efthimiou, Thomas Hanke, Ernst Thoutenhoofd, and Inge Zwitserlood, 44–49. Paris: ELRA.

Efthimiou, Eleni, Stavroulou-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, and Johanna Mesch 2020. *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*. Paris: ELRA.

Frishberg, Nancy. 1975. Arbitrariness and iconicity: Historical change in American Sign Language. *Language* 51:696–719.

Hanke, Thomas, M. González Rodriguez, and C. Paz Suarez Araujo. 2002. iLex—A tool for sign language lexicography and corpus analysis. Presented at the LREC 2002 Conference, Las Palmas de Gran Canaria, Spain, May 27–June 2.

Harris, Raychelle, Heidi M. Holmes, and Donna M. Mertens. 2009. Research ethics in sign language communities. *Sign Language Studies* 9 (2): 104–131. doi:10.1353/sls.0.0011.

Herrmann, Annika, and Markus Steinbach. 2011. Nonmanuals in sign languages. *Sign Language and Linguistics* 14 (1): 3–8. doi:10.1075/sll.14.1.02her.

Jayaprakash, Rekha, and Thomas Hanke. 2014. How to use depth sensors in sign language corpus recordings. In *Beyond the Manual Channel: 6th Workshop on the Representation and Processing of Sign Languages*, ed. Onno Crasborn, Eleni Efthimiou, Stravroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette H. Kristoffersen, and Johanna Mesch, 77–80. Paris: ELRA.

Johnston, Trevor. 2004. W(h)ither the deaf community? Population, genetics and the future of Auslan (Australian Sign Language). *American Annals of the Deaf* 148 (5): 358–375. doi:10.1353/aad.2004.0004.

Johnston, Trevor. 2008. Corpus linguistics and signed languages: No lemmata, no corpus. In *5th Workshop on the Representation and Processing of Signed Languages: Construction and Exploitation of Sign Language Corpora*, ed. O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd, and I. Zwitserlood, 82–87. Paris: ELRA.

Johnston, Trevor. 2009. Creating a corpus of Auslan within an Australian National Corpus. In *HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*, ed. Michael Haugh, Kate Burridge, Jean Mulder, and Pam Peters, 87–96. Somerville, MA: Cascadilla Proceedings Project.

Johnston, Trevor. 2016. *Auslan Corpus Annotation Guidelines,* February 2016 version. Macquarie University. http://www.auslan.org.au/about/annotations/.

Johnston, Trevor, and Adam Schembri. 2006. Issues in the creation of a digital archive of a signed language. In *Sustainable Data from Digital Fieldwork*, ed. L. Barwick and N. Thieberger, 7–16. Sydney: University of Sydney Press.

McKee, Rachel Locker, and Victoria Manning. 2015. Evaluating effects of language recognition on language rights and the vitality of New Zealand Sign Language. *Sign Language Studies* 15 (4): 473–497. doi:10.1353/sls.2015.0017.

Pollard, Robert Q., Jr. 1992. Cross-cultural ethics in the conduct of deafness research. *Rehabilitation Psychology* 37 (2): 87–101. doi:10.1037/h0079101.

Schembri, Adam, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, and Kearsy Cormier. 2013. Building the British Sign Language Corpus. *Language Documentation and Conservation* 7:136–154.

Singleton, Jenny L., Amber J. Martin, and Gary Morgan. 2015. Ethics, deaf-friendly research, and good practice when studying sign languages. In *Research Methods in Sign Language Studies: A Practical Guide*, ed. Eleni Orfanidou, Bencie Woll, and Gary Morgan, 7–20. West Sussex, UK: John Wiley and Sons.

Supalla, Ted, and Patricia Clark. 2014. *Sign Language Archeology*. Washington, DC: Gallaudet University Press.

Trippel, Thorsten. 2004. Metadata for time aligned corpora. In *Proceedings of the LREC 2004 Workshop: A Registry of Linguistic Data Categories within an Integrated Language Repository Area*, ed. Thierry Declerck, Nancy Ide, Key-Sun Choi, and Laurent Romary, 49–55. Paris: ELRA.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the LREC 2006 Conference*, ed. Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, 1556–1559. Paris: ELRA.

This is a section of doi:10.7551/mitpress/12200.001.0001

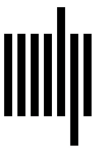
The Open Handbook of Linguistic Data Management

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 open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

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