D9.1 DEFINITION OF MINIMAL CONTENTS OF DATASET FOR PARTICIPATION

Work package	WP 9
Task	Task 9.1
Due date	30/06/2022
Submission date	29/06/2022
Deliverable lead	Radboud University
Version	1.0
Authors	Onno Crasborn, Hope Morgan (Radboud University)
Reviewers	Robin Ribback (STXT), Eleni Efthimiou (ATHENA)

Abstract	This report describes what types of language resources are needed for new languages to possibly make future use of technologies developed in EASIER.
Keywords	Language resources, under-resourced languages, sign language resources, corpus, lexicon

Grant Agreement No.: 101016982
Call: H2020-ICT-2020-2
Topic: ICT-57-2020
Type of action: RIA

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	23/06/2022	Document draft	Onno Crasborn, Hope Morgan (RU)
V0.2	28/06/2022	Internal Review 1	Robin Ribback (STXT)
V0.3	29/06/2022	Internal Review 2	Eleni Efthimiou (ATHENA)
V1.0	29/06/2022	Camera-ready submission	Onno Crasborn, Hope Morgan (RU)

DISCLAIMER

COPYRIGHT NOTICE

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable:	**Report**	
Dissemination Level		
PU	Public, fully open, e.g., web	✓
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to EASIER project and Commission Services	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

Funded by the Horizon 2020 Framework Programme of the European Union

EXECUTIVE SUMMARY

Language technologies are always developed initially based on large datasets. These are imperative to be able to develop innovations. For three-year projects like EASIER, that means working with sign languages for which such relatively large datasets are indeed available. What is needed for new languages to make use of the emerging technologies in the future likewise is datasets that are large enough for algorithms to work with. But what counts as large?

The present document aims to address this question for signed languages, from the context of the EASIER project. This project has a set of signed and spoken language pairings that it focuses on: BSL-English, DGS-German, DSGS-German, GSL-Greek, LIS-Italian, NGT-Dutch, and LSF-French. One intention of EASIER is to create technologies that will also lead to innovations for other combinations of signed and spoken languages. The present report focuses on signed languages, asking 1) what type of data are needed for new languages to become relevant for such technology development, bringing new technologies closer to the respective language communities, and 2) what size these datasets need to be. We thus hope to assist those (planning to be) involved in the development of sign language resources at present and in the near future.

At minimum, both corpus data and lexical data are required, but it is not straightforward to determine hard lower boundaries. This report advises to aim for the size of the present-day largest corpora, which is between 100-300 hours of primary (interaction) data. Annotating only a part of this data at the lexical level will already easily lead to a lexicon of 2500+ items. In addition to these quantitative benchmarks, the report also describes some of the dimensions of quality that are involved in data recording, annotation, and archiving, without going into great detail about the specific standards that might be desirable for each area. Implementation of each of these stages of documentation involves much detail, beyond the scope of this report. Therefore, key references to literature are provided throughout the report.

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE OF CONTENTS

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF TABLES

Funded by the Horizon 2020
Framework Programme of the European Union

1 INTRODUCTION

In 2020, the European Commission funded this research project in the domain of Information and Communication Technologies (ICT) aimed at technology for sign languages (call ICT-57 of Horizon 2020). This project runs for three full years from 2021-2023, and broadly aims to create innovations in the domain of machine-supported translation between signed languages and spoken languages. The availability of large datasets for the languages at hand is imperative to be able to innovate and develop such technologies. But what counts as large?

The present document aims to address this question for signed languages, from the context of the EASIER project. This project has a set of signed and spoken language pairings that it focuses on: BSL-English, DGS-German, DSGS-German, GSL-Greek, LIS-Italian, NGT-Dutch. The intention of this project is to create technologies that will also lead to innovations for other combinations of signed and spoken languages. The present report focuses on signed languages, asking 1) what type of data are needed for new languages to become relevant for such technology development, bringing new technologies closer to the respective language communities, and 2) what size these datasets need to be. We thus hope to assist those (planning to be) involved in the development of sign language resources at present and in the near future.

The two key types of language resources that are needed to document a dataset are a **lexicon** and an annotated **video corpus** of continuous sign language use (as in monologues, dialogues, etc.) that refers to the lexicon in its annotations. These two will be distinguished below, although some guidelines apply to both. While the present document focuses on the creation of sign language resources for technology development, most datasets tend to have multiple functions, addressing not only language documentation needs but also lexicography efforts and sometimes specific linguistic research questions.

At the end of this document, there is a section ‘Further Reading’ that includes references to literature that are not referred to in the text.

Funded by the Horizon 2020
Framework Programme of the European Union

2 APPROACH OF THE PRESENT REPORT

As the saying goes, for some people 'the only good data is more data'. There are various algorithms involved in machine translation between video-recorded sign language use and text or speech. This will result in higher-quality output when more data are available to train them on. For that reason, it is difficult to say what is the ideal or even the 'minimal' size for lexical and corpus resources. Rather, in this document we report on the collected experience of EASIER's project partners in creating sign language resources. These experiences have been gathered in the last 20+ years for most sign languages, shared in many workshops and strengthened through joint projects (e.g., the ECHO project and the resulting dataset[1]), and have followed new standards in video technology and set new standards for annotation of sign language data.

In the next section (section 3), we will therefore report on the current sizes of the datasets for the seven sign languages that are the focus of EASIER, allowing resource creators for other sign languages to see how their projects (current or proposed) can aim to reach the bandwidth represented by the datasets for these seven.

At the same time, *quality is as much of importance as quantity*; therefore, this document also highlights some important aspects of quality that will make new resources suited for emerging language technologies (section 4).

[1] https://www.ru.nl/cls/our-research/research-groups/sign-language-linguistics/completed-projects/completed-projects/echo/

Funded by the Horizon 2020 Framework Programme of the European Union

3 CURRENT SIZES OF SIGN LANGUAGE DATASETS AND AN ATTEMPT AT A GOOD TARGET SIZE

In the following two subsections, we first present information about the seven project languages of EASIER, followed by some other datasets for European sign languages that are outside the scope of the project. The data are taken from another EASIER report, “Overview of datasets for the sign languages of Europe” (Kopf et al. 2021), unless otherwise noted. Then in section 4, we will further discuss the content of datasets, which for some machine learning tasks can matter as much as the size.

3.1 DESCRIPTIVE STATISTICS ABOUT THE SEVEN EASIER SIGN LANGUAGES

In Table 1, the sizes of lexical and corpus datasets are specified in terms of the video source data and the extent to which these have been annotated. We singled out the two most important aspects of annotation: **glossing or tokenizing** (identifying all tokens of lexical types in a video) and **translation**. We will come back to corpus annotation in section 4. Although the table suggests homogeneity — and many similarities do exist — in fact corpus creators have approached these tasks in different ways (see section 4 for discussion). Details on each dataset can be found in Kopf et al. 2021, with source references listed there. As the many question marks make clear, not all information was easily accessible and included in Kopf et al.’s report.

In evaluating the size of a corpus, it is important to point out that for all seven languages, only a subset of the full corpus is annotated. For the size of the lexicon, we have tried to list the number of different sign types that occur in the corpus annotations, but most lexical datasets are in fact larger, containing items that have not yet been observed during the corpus annotation process (they may be in video data and not yet annotated, or are not yet in any video in the corpus).

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE 1: SIZE OF DATASETS FOR THE SEVEN SIGN LANGUAGES IN EASIER

Language	Corpus size (hours of video-recorded data)	Corpus size (number of signers)	Corpus size (hours of tokenized videos)	Corpus size (hours of translated videos)	Size of linked lexicon
DGS	560	330	90.9	90.9	14,064
NGT	72	92	15	15	3,300
BSL	125	249	18.4[2]	53.8[2]	<2,500
DSGS	–	–	–	–	–[3]
GSL (POLYTROPON corpus)	3,600 utterances	1	?	3,500 utterances	1,600
GSL (Dicta-Sign-GSL-v2)	5.5	16	5.5	5.5	?
LSF (CREAGEST corpus)	300	250	1	1	?
LSF (Dicta-Sign-LSF-v2)	11	16	11?	11	?
LIS	±100	180	?	0	?
Bandwidth across languages	5.5–560 hours	16–330 signers	5.5–90.9 hours	5.5–90.9 hours	Up to 14,064 lexical items

As the table shows, for some languages a single corpus is available (containing monologues and/or dialogues), for others a set of sentences has been recorded. Of the larger datasets, only a subset is annotated. Finally, typically the full documented lexicon is often larger than the lexicon needed to annotate the corpus data so far; or in other words, not all signs in the whole lexicon have yet been observed to occur in the corpus. The last column in table 1 only lists signs that have so far been attested in the corpus.

[2] Number calculated from the "Notes to v.3.0 release of BSL Corpus annotations: Translations", March 2017. Downloaded from https://bslcorpusproject.org/cava/, June 13, 2022.

[3] A large lexicon containing 9,000 signs is available for DSGS, but it has not been used for the annotation of any archived corpus.

Funded by the Horizon 2020 Framework Programme of the European Union

3.2 DESCRIPTIVE STATISTICS FOR SOME OTHER LARGE SCALE EU SIGN LANGUAGE DATASETS

In comparison, in the table below are some figures for other sign languages that have been relatively well documented.

TABLE 2: SIZES OF DATASETS FOR SOME OTHER EUROPEAN SIGN LANGUAGES

Language	**Corpus size** (hours of video-recorded data)	**Corpus size** (number of signers)	**Corpus size** (hours of tokenized videos)	**Corpus size** (hours of translated videos)	**Size of linked lexicon**
Irish SL	10	40	?	?	?
LSE	?	85	?	?	?
PJM	400	150	*505,000 tokens*	*10,000 clauses*	?
SSL	24	42	24	14	18,800?
VGT	140	120	–	?	–

Although information about the size of the annotations and associated lexicon is not always available, it is clear that these fall in the same wide bandwidth as the resources for the EASIER languages. For all languages, the sizes of the datasets are (very) small in comparison to what is available for speech and especially for text corpora, making all sign languages fall in the category of 'under-resourced languages.'

3.3 WHAT DOES THIS MEAN FOR YOU? RECOMMENDATIONS FOR THE SIZE OF NEW DATASETS

Projects like EASIER must deal with such variety as presented in section 3.1. While for some engineering tasks resources from different datasets and even different languages can be pooled, ideally tools for users would be able to handle the low number of resources for some languages. The quality of recognition, translation, and synthesis will of course increase for a language as more resources and larger datasets are available.

What this means for researchers that create language resources is that no hard rules can be given for the optimal size of new datasets. A balance will have to be found between investing in recording signs and corpus data (narratives, interaction data) and the time and budget spent on annotating those data. With the advent of machine-supported annotation around the corner, the approach taken for most current corpora — i.e., 'record more than you can annotate in a few years' — seems wise, and is also our recommendation. Around 100 hours of corpus data seems a good starting point, based on the datasets listed above.

If a lexicon is to be created from corpus annotation ('add a sign to the database whenever it is observed in the corpus during tokenisation'), the lexicon will quickly reach the size of thousands of lemmas during the annotation of the first few hours of data. If a lexicon is created before the

Funded by the Horizon 2020 Framework Programme of the European Union

process of corpus annotation (for instance as the basis of a dictionary), then a size of 2,500 lemmas for frequently-used concepts along with common function words will serve as a good basis for corpus annotation and synthesis – assuming that there is no regional or other variation. Adding variants for other regions may well multiply that dataset, depending on how much variation is present in the language (e.g., Stamp et al. 2014 for BSL).

Regional, age, or gender variation will also have an impact on how many signers need to be recorded; here, too, we see large variation in the above tables across datasets and languages. To create robust sign language recognition across different signing styles, a corpus with many different signers will be mandatory. For other types of applications, like providing example sentences for dictionary entries, just a single signer might be recorded. So, the purpose of the dataset will influence the number of signers that will be recorded. The size of the country or region, the number of subcommunities in that country, and the related degree of variation across (sub)regions are most likely to affect the number of signers, too. The larger corpora in Europe have strived for recording at least 5-10 men and women in each region.

In the remaining sections, we discuss the more qualitative aspects of resource creation, including what is recorded and how, and some issues in corpus annotation.

Funded by the Horizon 2020
Framework Programme of the European Union

4 QUALITATIVE RECOMMENDATIONS

The recommendations in this section in part reflect current technological possibilities and in part reflect what corpus projects for other sign languages have been doing. This does not exclude doing things in a different way. It is also important to point out that because this report is focused on data benchmarks, it does not include a separate section on the many ethical considerations in language documentation, language resource collection, and the use of these resources – although these are mentioned where relevant. Yet this is a crucial aspect of any sign language documentation project. Collaborating with the deaf community and the inclusion of deaf colleagues at each step is now an expected part of the process of documenting a sign language. There are several excellent discussions of this in the literature by now; see e.g. Harris et al. 2009, Hochgesang 2015, Singleton et al. 2015, and De Meulder 2021.

4.1 CONTENT OF CORPUS DATA

Whereas text and speech corpora are often harvested from available sources like archived or online media, sign language corpora almost always contain data recorded for the purpose of language documentation. Therefore, most corpora control the balance of participants in terms of sociolinguistic properties that are relevant for that language community. Typically, this means including participants from different age groups, regions within the community, and different genders. Other factors may be of relevance, and working together with the language community is vital as in any documentation work to bring these factors to the fore.

Sign language research has typically focused on Deaf native signers: people who feel that they are part of the deaf community in question and are considered by others to be, and who grew up acquiring the language from an early age. As few deaf people acquire sign language from birth from deaf family members, corpora have often broadened their criteria to include 'early learners', with varying cut-off ages for the start of sign language exposure. The whole language community arguably also includes hearing children of deaf adults and hearing siblings of deaf children, who are often native signers, as well as many late learners, whether hearing, hard of hearing, or deaf.[4] While current corpora have not incorporated this level of sociolinguistic diversity, it is a direction that future corpus collectors are invited to consider.

Almost all present corpora consist of a mix of tasks, from free conversation to elicited discussion to narratives to sign or sentence elicitation – and much more.[5] For an inventory, see Kopf et al. 2021. A good balance between the various types of tasks (incl. free conversation) is desirable. Whereas the more controlled tasks allow for a good comparison of all signers (and thus the sociolinguistic variables they represent), spontaneous interaction is most representative of everyday sign language use. Language technologies will ultimately need to be able to deal with a variety of types of language use, so variety of tasks in a corpus will be important for technological innovation and the adoption of existing technologies for new languages.

To create a useful and representative corpus it is important to have comparable tasks across signers, but also to ensure that recordings are made in not too long a timespan. If the corpus

[4] There are also other issues with prioritizing and idealizing a "native speaker/signer"; see Cheng et al. 2021 for a recent discussion.

[5] One of the few corpora that only contains spontaneous interaction in natural settings is the NGT Interactive Corpus (https://hdl.handle.net/1839/00-0000-0000-0021-8357-B). A corpus like that will be ideal for linguistic studies of spontaneous interaction but may be harder to parse by sign recognition algorithms because of the larger variety in camera positions, for instance.

Funded by the Horizon 2020 Framework Programme of the European Union

aims to be a snapshot of the language as used at time X, that time should be a window of maximally a few years. Language change and cultures change, and the closer together the recording dates are, the more coherent the corpus will be.

Finally, the corpora in section three all contain monologue or, more typically, dialogue settings. These are easier to record and annotate than multi-party interactions with three or more signers, and for that reason are a wise first choice for a first corpus for a language. Likewise, and implicit in what was written above, interactions between native and non-native signers or deaf and hearing signers are often not the choice for a first corpus documenting the use of a sign language, although they may be equally ecologically valid.

4.2 TECHNICAL SPECIFICATIONS FOR VIDEO CORPORA

Any project will have to manage its own balance between costs, quality, and size, and it is not always necessary to have the highest possible technical quality. Also, continuing advances in technology can change the balance of factors. At the same time, it is important to create recordings that can serve a variety of purposes and meet certain criteria. Starting with recording video with one or more cameras, quality can be specified in terms of spatial resolution, temporal resolution, and shutter speed. **Spatial resolutions** of 1280x720 ('half HD'; HD = high definition) and 1920x1080 pixels ('full HD') are now the lowest that even simple cameras can record. Full HD is a safe lower boundary that will satisfy most needs, even if 4K and in the future 8K are becoming the new high-end standards (standing for four respectively times HD). The **temporal resolution** is expressed in the number of frames per second. Current cameras can all record full frames (rather than alternating frames with even and odd lines, called interlacing), which is highly desirable to record body motion. International standards differ, European frame rates are typically 25 or 50 frames per second; more expensive cameras can record 100 or more frames per second. All are fine in principle, if the **shutter speed** is high enough to avoid blurred motion.[6] When recording at 50 frames per second, a shutter speed is often the default. Many cheaper cameras do allow to select combinations of spatial resolution and frame rate, but do not allow to manually adjust the shutter speed. **Good lighting** is especially important for those cameras (including those in smartphones): when more light comes into the camera, the automatic shutter speed will be faster, and the less motion blur will result.

In addition, some issues about the filming environment should be mentioned. First, it is important that there are no strong shadows on the face. Soft lighting from dedicated film lights is always cheap in comparison to the investment in cameras, so we recommend investing in that. Second, the choice of clothing and the combination of skin colour and clothing is also worthwhile to pay attention to: at best, participants are wearing even-coloured clothes without any stripes or other patterns on it, and which contrast a bit with their skin colour. In that way, it will be easier for computer vision algorithms to correctly track the movements of the hands and fingers.

Existing corpora have made different choices for equipment and video settings, partly based on the available budget. It is generally advisable to have at least one camera focused on each participant, and one on the whole scene that can also record interactions with the moderator. Some corpora have however also used special cameras for top views and/or face views. Multiple cameras can be synchronised by importing a shared time code, or by a post-hoc synchronisation based on a sound like hand clapping that all cameras record at the start. Next to plain video cameras, some datasets have been created with three-dimensional (3D) video

[6] There are many online resources explaining the relation between shutter speed, frame rate, and motion blur. For instance, see https://exposureworks.co.uk/video-frame-rates-shutter-speed-and-motion-blur/.

Funded by the Horizon 2020 Framework Programme of the European Union

cameras, infrared depth cameras or motion capture systems. We leave these out of consideration here, but refer to the *sign-lang@LREC anthology* for descriptions of various datasets.

Finally, recorded and edited video needs to be compressed to an archive format, that will in part depend on the purpose. While for some computer processing tasks, maximal temporal and spatial resolution with limited compression may be desirable, but for annotating linguistic data in ELAN, a lower resolution may be sufficient. (See also below under archiving, section 4.5.)

4.3 ANNOTATION OF CORPUS DATA

4.3.1 Glossing of manual lexical signs

A corpus is not a corpus until it is annotated and searchable, whether by humans or by automated routines. As the tables in section 3 already suggested, sentence-level translation and glosses for manual signs are the most common annotation layers added to video data. Even for these, conventions differ. And even for just these two, annotation is a manual process that takes a lot of time. Estimates of 250 times real time (250 hours to annotate one hour of dialogue) are not uncommon. Therefore, no sign language corpus has yet been fully annotated, and even if a large budget is available, it is likely that there will be more video data than one can annotate. For developing automated recognition and translation software, even the largest set of annotations currently available is presently too small to allow for top-notch machine translation of any signed video. So, while for language documentation purposes a focus on recording a lot of language varieties is a priority, for language technology, annotation should be a priority.

None of the sign corpora involved in EASIER have attempted to create a full phonetic transcription of all relevant signals in communication. Phonetic transcription (esp. of non-manuals) is an even more time-consuming task. The focus has rather been on translating sentences (or equivalent or larger units) to a spoken language, and on the ID-glossing (tokenisation) of all manual activity. For some corpora like the DGS corpus, ID-glossing goes hand in hand with a phonetic transcription in HamNoSys (of that manual activity and any consistent non-manual activity that comes with the sign), for others (BSL, NGT) this is not the case – although the phonological form of signs is always encoded in some manner. In addition to translation and ID-glossing, the (orthographic) transcription of mouth actions is also not uncommon, and may be of value for machine processing.

The notion of an **ID-gloss** as the mediating text string between corpus and lexicon, developed over the last ten years, starting with Johnston (2010), is a crucial one. An ID-gloss is not a semantic label for a sign, but it refers to a specific form in the lexicon. This reference could take any form, but for human readability, it is most often a spoken language word that best approximates the most general meaning of a sign. It must be unique, in that a single gloss identifier consistently refers to the same sign form in the lexicon, whether at the level of the lemma, the full (inflected) form, or any other level, if it is consistently done.

This still leaves open many aspects of lexical signs and morphological complexity, like compounding and numeral incorporation; we refer to the annotation conventions in the further reading section for more discussion of these.

4.3.2 Other aspects of manual signing

Funded by the Horizon 2020 Framework Programme of the European Union

There are some manual activities in natural signing that may not constitute full (inflected or uninflected) signs that can be listed in the lexicon, but are 'partly lexical' and partly gestural.[7] Pointing actions, for instance, have a specific handshape and movement direction that can be specified in the lexicon, while the direction and/or location of pointing are determined by the spatial context of the discourse and/or the visible surroundings. Something similar holds for classifier constructions, where the handshape (and sometimes orientation) may be lexically specified, whereas the location and movement (and sometimes orientation) are dependent on the spatial (discourse) context.

Similarly, different approaches have been taken with regards to the use of left-handed vs. right-handed signs (both in left-handed signers and in dominance reversal within the discourse) and the long holds of one hand while the other hand continues to sign (sometimes called 'buoys').

Corpora have dealt with all the above issues in different ways. The EASIER report, *Specification for the harmonization of sign language lexicons* (Kopf et al. 2022) provides a detailed description about the differences and similarities of these annotations (see also Cormier et al. 2016). For EASIER, a standard has yet to be formalised.

4.4 LEXICON

As the repository for all fully lexical material and sometimes the lexical part of partly-lexical activities of the hands, a digital lexicon is important as a tool for consistent corpus annotation. It also plays an important part in language synthesis (animated signing by avatars). Sometimes, a lexicon also serves as a public dictionary (learner or reference). Depending on the goal, the desired minimal size will differ. While corpus annotation can happen offline (with tools like ELAN) or online (with an integrated database like iLex that hosts both lexicon and annotations), lexical databases nowadays all tend to be online. This implies that a **server infrastructure** will be necessary, with sufficient technical support for maintaining this infrastructure.

As a basic list of concepts with an ID-gloss, a lexicon should at least be able to represent the form and the meaning of each sign. In its most simple form, the meaning is often a list of translation equivalents in a spoken language. The representation of the form can take the shape of a phonetic transcription in HamNoSys, or a list of values for a set of fields that characterise basic properties of signs. Here, too, there is no standard, and different databases have used different approaches to encoding the phonology (see Kopf et al. 2022). Some are richer (allowing to differentiate each form from all other forms), some are leaner (encoding basic handshape, movement, and location properties for the two hands). For the purpose of sign synthesis, a rich encoding may be needed, but for the lookup of signs during annotation, a 'quick and dirty' description in terms of Stokoe's main sign parameters would suffice. Here, too, no conventions have yet been specified for EASIER.

A matter of special interest needs to be **lemmatisation**: how to handle variants in form, whether systematic and morphological (such as inflections) or sociolinguistically conditioned or simply of an unknown background? With the basic premise that *every form needs to have its own ID-gloss*, there will still be basic discussions on what needs to be classified as instances of the same lemma, and what counts as different. Different approaches are possible, and it is beyond the scope of this document to describe them here; see Fenlon et al. 2015 for BSL, Konrad 2011 and Langer et al. 2016 for DGS. However, using an existing database platform

[7] There is much discussion among sign linguists about signs that come under this overall description. Regardless of how they are characterized, however, these types of signs have properties that necessitate a different treatment in annotation than "fully lexicalized" signs.

Funded by the Horizon 2020 Framework Programme of the European Union

(like iLex or one of the Signbanks, see Cassidy et al. 2018) in conjunction with their guidelines will create consistency in lemmatisation, which is essential.

4.5 ARCHIVING

Even if data are made accessible via a server that you maintain yourself, it is still advised to **archive a copy** at a larger language archive like The Language Archive of the Max Planck Institute for Psycholinguistics[8] or that of ELRA[9] or ELAR[10] (both also in Europe). Using standard **metadata** as required by such archives will help make data accessible to search engines and larger repositories of linguistic data. The archive can then be trusted to ensure availability of data for a longer timespan than what most research groups or universities can guarantee. That said, it may still be worthwhile maintaining an iLex or Signbank **server** for immediate research needs. If the hosting institute can guarantee long-term availability, this may even suffice as an archive for future generations.

More generally, published data should be findable, accessible, interoperable, and re-usable ('**FAIR**', see https://www.go-fair.org/fair-principles/). That is, a dataset will have most impact if it follows international standards where available, and if the publication is not an afterthought but a goal. See Schulder & Hanke 2022 for discussion.

Data management in the broader sense — i.e., taking good care of the whole lifecycle of your data — has received increasing attention and is now considered part of the culture of open science that current technologies facilitate. A recent handbook on linguistic data management also includes chapters on sign language data (Berez-Kroeker et al. 2022) and is highly recommended.

[8] https://archive.mpi.nl

[9] http://www.elra.info/en/services-around-lrs/distribution/

[10] https://www.elararchive.org

Funded by the Horizon 2020 Framework Programme of the European Union

5 FURTHER READING

Aside from the cited references listed in the next section, much can be learned from the short papers in the **sign-lang@LREC Anthology** maintained by the University of Hamburg: the bi-annual Language Resources and Evaluation Conference (LREC) has always had a large contribution from the sign language community. The papers in the anthology concern both the creation of language resources as well as their use for the development of sign language technologies. See https://www.sign-lang.uni-hamburg.de/lrec/.

5.1 ANNOTATION CONVENTIONS FOR SOME CORPORA

- Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., & Regen, A. (2019). *Public DGS Corpus: Annotation Conventions*. Hamburg University. https://www.sign-lang.uni-hamburg.de/dgs-korpus/arbeitspapiere/DGS-Korpus_AP03-2018-01v02_en.pdf
- Johnston, T. (2019). *Auslan Corpus Annotation Guidelines.* https://auslan.org.au/about/annotations/
- Cormier, K., & Fenlon, J. (2014). *BSL Corpus Annotation Guidelines*. https://bslcorpusproject.org/wp-content/uploads/BSLCorpusAnnotationGuidelines_23October2014.pdf
- Crasborn, O., Bank, R., Zwitserlood, I., Kooij, E. van der, Ormel, E., Meijer, A. de, & Sáfár, A. (2015). *Annotation conventions for the Corpus NGT. Version 3*.

5.2 SOME OTHER USEFUL PUBLICATIONS, ESPECIALLY ON ANNOTATION

- Crasborn, O. (2015). Transcription and notation methods. In E. Orfanidou, B. Woll, & G. Morgan (Eds.), *Research methods in sign language studies. A practical guide* (pp. 74–88). John Wiley & Sons.
- Frishberg, N., Hoiting, N., & Slobin, D. I. (2012). Transcription. In R. Pfau, M. Steinbach, & B. Woll (Eds.), *Sign Language. An international handbook* (pp. 1045–1075). De Gruyter Mouton.
- Hodge, G., & Crasborn, O. (2022, in press). Good practices in annotation. In J. Hochgesang & J. Fenlon (Eds.), *Sign language corpora*. Gallaudet University Press.
- Konrad, R., & Langer, G. (2009). *Synergies between transcription and lexical database building: The case of German Sign Language (DGS)*. Proceedings of the Corpus Linguistics Conference (CL2009). University of Liverpool, UK, 20-23 July 2009.
- Miller, C. (2006). Sign Language: Transcription, notation, and writing. In K. Brown (Ed.), *Encclopedia of language and linguistics* (Vol. 8, pp. 353–354). Elsevier.
- Ochs, E. (2006). Transcription as theory. In A. Jaworski & N. Coupland (Eds.), *The discourse reader* (pp. 166–178). Routledge.
- Crasborn, O., & Sáfár, A. (2016). An annotation scheme to investigate the form and function of hand dominance in the Corpus NGT. In M. Steinbach, R. Pfau, & A. Herrmann (Eds.), *A Matter of Complexity: Subordination in Sign Languages* (pp. 231–251). Mouton de Gruyter.

Funded by the Horizon 2020 Framework Programme of the European Union

REFERENCES

[1] Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E., & Johnston, T. (2018). Signbank: Software to Support Web Based Dictionaries of Sign Language. In 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 (Ed.), *Proceedings of LREC 2018* (pp. 2359–2364). ELRA.

[2] Cheng, L.S., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A. and Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.715843

[3] Cormier, K., Crasborn, O., & Bank, R. (2016). *Digging into Signs: Emerging Annotation Standards for Sign Language Corpora* (E. Efthimiou et al, Ed.; pp. 35–40). ELRA.

[4] De Meulder, M. (2021). Is “good enough” good enough? Ethical and responsible development of sign language technologies. *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, 12–22. https://aclanthology.org/2021.mtsummit-at4ssl.2

[5] Harris, R., Holmes, H. M., & Mertens, D. M. (2009). Research Ethics in Sign Language Communities. *Sign Language Studies*, *9*(2), 104–131. https://doi.org/10.1353/sls.0.0011

[6] Hochgesang, J. A. (2015). Ethics of researching signed languages: The case of Kenyan Sign Language. In A. C. Cooper & K. K. Rashid (Eds.), *Citizenship, politics, difference. Perspectives from sub-Saharan signed language communities* (pp. 9–28). Gallaudet University Press.

[7] Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus LInguistics*, *15*(1), 104–129.

[8] Kopf, Maria, Schulder, Marc, & Hanke, Thomas. (2021). *Overview of Datasets for the Sign Languages of Europe*. Deliverable 6.1, EASIER Project. Hamburg: University of Hamburg. https://doi.org/10.25592/UHHFDM.9560]

[9] Kopf, Maria, Schulder, Marc, & Hanke, Thomas. (2021). *Specification for the Harmonization of Sign Language Lexicons*. Deliverable 6.2, EASIER Project. Hamburg: University of Hamburg. https://doi.org/10.25592/uhhfdm.9841

[10] Schulder, M., & Hanke, T. (2022). How to be FAIR when you CARE: The DGS Corpus as a Case Study of Open Science Resources for Minority Languages. *Proceedings of LREC 2022*, 164–173.

[11] Singleton, J. L., Martin, A. J., & Morgan, G. (2015). Ethics, deaf-friendly research, and good practice when studying sign languages. In E. Orfanidou, B. Woll, & G. Morgan (Eds.), *Research methods in sign language studies. A practical guide* (pp. 7–20). John Wiley & Sons.

[12] Stamp, R., Schembri, A., Fenlon, J., Rentelis, R., Woll, B., & Cormier, K. (2014). Lexical Variation and Change in British Sign Language. *PLoS ONE*, *9*(4).

Funded by the Horizon 2020 Framework Programme of the European Union