Proceedings of the 10th Workshop on the Representation and Processing of Sign Languages (sign-lang@LREC 2022), pages 102–109
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources

Maria Kopf, Marc Schulder, Thomas Hanke
Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Germany
{maria.kopf, marc.schulder, thomas.hanke}@uni-hamburg.de

Abstract

One of the challenges that sign language researchers face is the identification of suitable language datasets, particularly for cross-lingual studies. There is no single source of information on what sign language corpora and lexical resources exist or how they compare. Instead, they have to be found through extensive literature review or word-of-mouth. The amount of information available on individual datasets can also vary widely and may be distributed across different publications, data repositories and (potentially defunct) project websites.

This article introduces the *Sign Language Dataset Compendium*, an extensive overview of linguistic resources for sign languages. It covers existing corpora and lexical resources, as well as commonly used data collection tasks. Special attention is paid to covering resources for many different languages from around the globe. All information is provided in a standardised format to make entries comparable, but kept flexible enough to allow for differences in content. The compendium is intended as a growing resource that will be updated regularly.

Keywords: Survey, Sign Languages, Corpora, Lexical Resources, Metadata

1. Introduction

Recent decades have seen a marked increase in the creation of digital resources for signed languages. This has opened up new possibilities for data driven research, such as computational and corpus linguistics, including work involving multiple languages or resources. Identifying and comparing suitable resources can still be a challenging task, however, requiring extensive literature review, web search and the use of personal contacts. The amount of information available on individual datasets can vary widely and may be distributed across different publications, data repositories and (potentially defunct) project websites. Documentation may also exist only in the local language(s) of the dataset creators, introducing additional barriers to international research. Once the information is gathered, different datasets may still be difficult to compare, as even basic meta-information like the size of a dataset may be reported in a variety of ways. These hurdles harm linguistic diversity as they discourage studies across multiple resources and favour the use of only the most well-known datasets.

To support researchers in their work, this paper introduces the **Sign Language Dataset Compendium**, an extensive overview of available digital datasets of signed languages. It covers both **corpora** and **lexical resources**, providing structured information and metadata, literature references, and pointers to where the data or more information can be obtained. It also provides an index of commonly used **corpus data collection tasks** to assist researchers in finding corpora with comparable contents. Additional topics, such as a documentation of existing annotation conventions, may be added in the future.

The entry for each corpus, lexical resource and collection task consists of the following elements:

1. a free-form description;
2. a structured info table;
3. corpus-specific task information (if applicable);
4. a list of references.

The core of each entry is the info table, which structures information using thematic categories commonly applicable to the given type of resource, such as size, linguistic information, participant demographic, data formats, licence conditions and more. The category fields follow a regular pattern, but there is enough flexibility to allow for differences in content.

The aim of the compendium is to help researchers find data that represent each language as it is used naturally by signers with L1 language proficiency. Corpora should contain (semi-)spontaneous language production rather than prepared utterances or translations of spoken language content. As such it does not cover interpreted television broadcasts or language acquisition datasets.

The compendium is available both as a website[1] and as a static document[2]. At the time of writing it describes 40 corpora, 63 lexical resources and 27 data collection tasks, covering 72 different sign languages. The compendium is intended as a growing resource that will be updated regularly.

[1] https://www.sign-lang.uni-hamburg.de/lr/compendium/

[2] https://doi.org/10.25592/uhhfdm.10210

2. Background

When looking for linguistic resources, three kinds of centralised information sources can be of relevance: curated lists, metadata repositories and data repositories. Each of these fulfil separate, if overlapping, important functions in the language data ecosystem. They will be described in the following subsections.

2.1. Curated Lists

Curated lists are resource descriptions compiled by a single author or editorial group. Commonly they strive to provide a comprehensive overview of resources that lie within their chosen scope. They describe the resources and often specify where they can be found, but do not store or host the data themselves. The compendium introduced in this article is such a curated list.

There exist a number of curated lists on sign language resources. As they were created for a number of different purposes, the type of information they provide differs considerably, as does the selection of resources and languages considered. While some are being maintained, others are snapshots of a specific point in time.

Originally created for the appendix of his dissertation, Konrad (2012) provides a detailed tabular overview of 34 sign language resources, identifying various linguistic properties of each resource. In her journal article, Schmaling (2012) provides a detailed overview of dictionaries for African sign languages. She focuses on print-media dictionaries, but also describes two resources providing video materials. Hartzell (2022) created an informal compilation of language resources for minority languages in Egypt, including eight resources for Egyptian Sign Language. As part of their overview website on automatic sign language processing, Moryossef and Goldberg (2021) include a table of sign language resources they consider suitable for such tasks. At the time of writing the table covers 36 resources and provides brief information regarding their size, licence, primary reference and data location. The website of the African Sign Language Resource Center[3] provides information on sign languages used in African countries. While the website is still under active development, it already contains profiles for 54 countries, offering general information on their deaf populations and used sign languages. In several cases, the profiles identify existing language resources, although not necessarily where to find them.

2.2. Metadata Repositories

Like curated lists, metadata repositories collect information about datasets without hosting the datasets themselves. Unlike curated lists, the required metadata is usually provided by the dataset creators, either through submission forms or harvested from metadata files that the creators host at dedicated locations. These metadata files must match supported formats such as Dublin Core, OLAC or CMDI. Depending on the metadata format, the design of the repository and the amount of information provided by creators, entries may provide only generic dataset information, general language dataset information or even information specific to sign language datasets. In the following we describe notable repositories that were designed specifically for language data and include entries for sign language resources.

The virtual library of the Open Language Archives Community[4](OLAC) harvests metadata from a number of participating language archives (Simons and Bird, 2003) using the OLAC metadata standard for language data (Bird and Simons, 2001).

Similarly, the CLARIN Virtual Language Observatory[5] (VLO) collects information on language resources, tools and services (Van Uytvanck et al., 2012; Goosen and Eckart, 2014). Metadata is harvested from various CLARIN centers and a small number of other providers. It supports multiple metadata standards and can represent datasets as hierarchical structures, allowing the interlinking of dataset collections, subcollections and individual components.

The LRE Map[6] (Calzolari et al., 2010) by the European Language Resources Association (ELRA) follows a different approach. Information is collected as part of the article submission process of participating conferences and workshops. Authors publishing articles about new or updated datasets are requested to fill out a metadata form for each. The forms are kept short to encourage many authors to fill them out, so they cover fewer aspects than VLO and OLAC.

2.3. Data Repositories

Data repositories host and archive datasets and provide them together with their metadata. Often submission of data is restricted to participating organisations. Depending on the focus of the repository, the metadata standards it uses and the information provided by data creators, resource descriptions may be more or less detailed. Unlike curated lists and metadata repositories, data repositories focus on representing the data they themselves host, rather than giving a general overview of available data.

Among the data repositories tailored specifically for language resources are The Language Archive[7] (TLA), the Endangered Languages Archive[8] (ELAR), META-SHARE[9] and the commercial ELRA Catalogue[10]. TLA and ELAR are noteworthy for explicitly taking sign languages into account in their categorisation and metadata structures. Each of these repositories lists several sign language datasets.

[3] https://africansignlanguagesresourcecenter.com

[4] http://www.language-archives.org

[5] https://vlo.clarin.eu

[6] https://lremap.elra.info

[7] https://archive.mpi.nl/tla

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

[9] http://www.meta-share.org

[10] https://catalog.elra.info

3. Creating the Compendium

The Sign Language Dataset Compendium originally evolved out of two other information collection efforts led by the authors of this article. The first was the creation of the *sign-lang@LREC Anthology*[11] (Hanke et al., 2021), the proceedings archive of the *Workshop Series on the Representation and Processing of Sign Languages*. To enrich the archive with additional metadata, a literature review of its 363 publications was performed to determine which datasets and tools each article introduced or used, which languages it addressed and which project it was part of. For each of these categories, an index was created and each entry enriched with basic information, such as its licence, links to the resource or project, or language identifiers.

The second effort was the *Overview of Datasets for the Sign Languages of Europe* (Kopf et al., 2021), a public project deliverable for the EU project EASIER[12]. This expanded review of literature, dataset and project websites and personal correspondence with data creators resulted in a structured report on 67 datasets (26 corpora and 41 lexical resources) and 26 data collection tasks, covering 24 languages. Since the EASIER project aims to develop machine translation technologies for signed and spoken languages of the European Union, the report focused on resources for European sign languages that were suitable for such tasks.

While both efforts fulfilled their set goals, the limitations of their scope meant that neither could function as a general global overview of sign language datasets. To fill this gap we decided to create the compendium, which would cover resources from across the entire globe. Naturally, this increase in scale also introduced new questions regarding curation criteria (see Section 3.1), in what format the compendium should be released (see Section 3.2) and how to best summarise information (see Section 3.3).

3.1. Curation

When choosing which resources to include in the compendium, a balance must be struck between quantity and quality. On the one hand it is our goal to provide a comprehensive overview of resources for as many languages as possible, on the other hand we wish to focus on resources that can be of use to the core audience of the compendium, corpus linguists and computational linguists. Curation criteria help define which resources should be included, but also which resources should be prioritised as we work on expanding the compendium.

The starting point of the compendium are the resource descriptions of Kopf et al. (2021). However, the curation criteria for that report were designed for the European resource landscape and needs of machine translation research. For the purposes of the compendium they had to be revised.

In selecting suitable curation criteria, we had to take into account that there exist strong imbalances between languages in the size and number of available resources. To address this we chose a two-tiered approach of minimum and strict requirements. All resources must meet the minimum requirements, but if some resources for a given language also meet the strict requirements, other resources for that language are not (yet) listed.[13] The conditions are applied to corpora and lexical resources separately, so a language can be subject to strict conditions for one and minimum restrictions for the other. This regulates the number of included resources for comparatively well-resourced languages without disqualifying less-resourced languages entirely.

The curation criteria for the compendium are as follows:

General criteria for resources

1. *Must include video data*: Motion is an essential part of sign languages; still images and drawings alone are not sufficient.

2. *No sign-supported systems*: The compendium covers only sign languages, systems to support spoken language with signs are not included.

3. *No language acquisition data*: Language acquisition research is a specialised area of linguistics with different data requirements than post-acquisition research. Consequently, descriptions of acquisition datasets require a different focus, which would require an extension of the compendium structures. For the time being, such an extension is outside the scope of the compendium.

4. *No historical sign languages*: Similar to language acquisition data, data about historical languages is outside the scope of the compendium in its current phase.

5. *Data must be attainable*: There needs to be a clearly defined way of accessing the resource. This may for example be a download location or a point of contact. A resource is not included if access by third parties is generally ruled out or if it is not available for other reasons, such as a lack of points of contact or storage and file formats that can not be accessed by current computer systems.

Corpora

6. *Must be (semi-)spontaneous signing*: The corpus should predominantly represent natural use of language, rather than prepared, interpreted or translated utterances.

7. *L1 signers*: The participants should be L1 users of the language.

[11] www.sign-lang.uni-hamburg.de/lrec/

[12] www.project-easier.eu/

[13] This limitation may be revisited in the future, after sufficient coverage across languages has been achieved.

8. *Annotation*: The minimal requirements for a sign language corpus to be machine readable are a free translation and ID-glosses (Johnston, 2010). Therefore corpora must at least have a partial translation and/or gloss annotation.

9. *Size*: Monolingual corpora must include at least 5 hours (minimum) or 10 hours (strict) of sign language recordings. Multilingual corpora are exempt.

Lexical Resources

10. *Must include index*: Individual lexemes must be directly accessible through an index, e. g. of glosses, translational equivalents or phonetic description. This excludes datasets that collect many lexemes in a single recording without identifying the starting timestamps of the discrete entries.

11. *Size*: Lexical resources must cover at least 100 (minimum) or 1000 (strict) different signs. Multilingual corpora are exempt.

Data Collection Tasks

12. *Used by multiple resources*: Collection tasks are included if they were used in the creation of more than one of the corpora described by the compendium.

Developing the described criteria was an organic process that went hand in hand with the inspection of potential resources. They may be adjusted further as the compendium grows over time.

3.2. Publication Formats

The compendium will be published in two formats: As a static report and as a website.

The report is provided as a PDF document, structured similarly to Kopf et al. (2021). As such it can be used offline or printed out and individual versions are easily cited and archived. Each version is registered with its own unique persistent identifier.

The website provides dynamic access to information by making it browsable through various indices and filters. For example, in the language index, each language provides a list of all resources that contain it. To make it easier to find the correct language in the index, a text filter allows users to search for it by its various names, acronyms and identifiers (see Section 4.6). More filters will be added in future releases, as the compendium is developed further.

3.3. A Descriptive Approach to Standardisation

A central goal of the compendium is to present information in a standardised structured format that makes it easy to inspect and compare entries. Dataset factors such as size, licence or data format and linguistic information like participant demographic or annotated phenomena should always be described the same way. In practice, this proved to be a complex challenge, both due to the complexity of language resources and the varied availability of documentation. Corpus size, for example, might be specified in terms of recorded hours, number of transcribed tokens/types, file size or number of files. Even within these categories, differences could be observed in what values were reported, e. g. whether *recorded hours* counted individual camera angles as separate recordings or the same time span. For linguistic information, this variability was even more pronounced, due to the varying goals of different resources and the large variety of annotation practices in the sign language research community (cf. Kopf et al. (2022))

To address this challenge, a descriptive approach was chosen. It was started in Kopf et al. (2021) and continued for the compendium. Information for a variety of resources was gathered first and based on what information could consistently be determined for most resources, categories were defined. Within each category, descriptions were kept free-form to allow suitable documentation of each resource, although as patterns emerged, a consistent format and terminology was employed where appropriate.

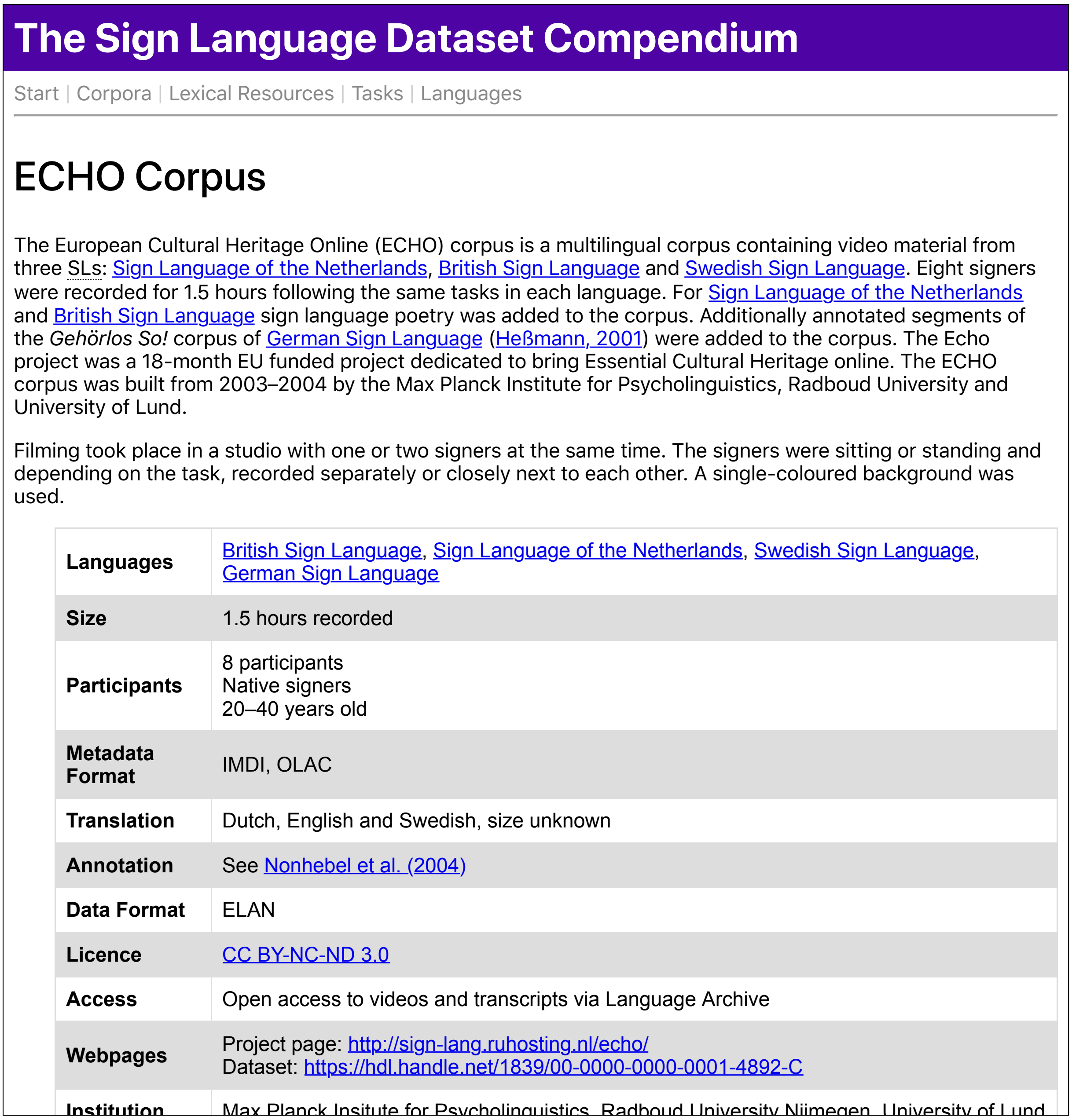
The advantage of this approach is that it ensures that individual entries are not restricted by pre-defined vocabularies and categories. Its downside is that, at this stage, it does not integrate with machine-readable metadata standards and closed vocabularies as they are recommended for modern open science practices such as the FAIR principles (Wilkinson et al., 2016). However, to enable its multiple output formats (see Section 3.2) the internal formatting of compendium entries already uses a set of semantic XML tags. This tag inventory will be further extended in future to allow the extraction of machine-readable information without harming the flexibility of human-readable contents.

4. Compendium Content

The compendium is intended as a resource overview for digital sign language resources. It collects two types of datasets: corpora and lexical resources. In addition to this it compiles information on data collection tasks commonly used in the creation of different corpora. It also provides basic entries for each language.

The information provided in the compendium is compiled from public resource documentation, research articles, inspection of public data and personal correspondence with resource creators.

Each compendium entry consists of a free-form text description, a structured info table and a list of references. The categories of the info table are described in the following subsections. There are categories specific to corpora (Section 4.1), to lexical resources (Section 4.2), general dataset categories applicable to both (Section 4.3) and categories for data collection tasks (Section 4.4). In addition, corpora and tasks contain tables providing information specific to individual



The Sign Language Dataset Compendium

ECHO Corpus

The European Cultural Heritage Online (ECHO) corpus is a multilingual corpus containing video material from three SLs: Sign Language of the Netherlands, British Sign Language and Swedish Sign Language. Eight signers were recorded for 1.5 hours following the same tasks in each language. For Sign Language of the Netherlands and British Sign Language sign language poetry was added to the corpus. Additionally annotated segments of the *Gehörlos So!* corpus of German Sign Language (Heßmann, 2001) were added to the corpus. The Echo project was a 18-month EU funded project dedicated to bring Essential Cultural Heritage online. The ECHO corpus was built from 2003–2004 by the Max Planck Institute for Psycholinguistics, Radboud University and University of Lund.

Filming took place in a studio with one or two signers at the same time. The signers were sitting or standing and depending on the task, recorded separately or closely next to each other. A single-coloured background was used.

Languages	British Sign Language, Sign Language of the Netherlands, Swedish Sign Language, German Sign Language
Size	1.5 hours recorded
Participants	8 participants Native signers 20–40 years old
Metadata Format	IMDI, OLAC
Translation	Dutch, English and Swedish, size unknown
Annotation	See Nonhebel et al. (2004)
Data Format	ELAN
Licence	CC BY-NC-ND 3.0
Access	Open access to videos and transcripts via Language Archive
Webpages	Project page: http://sign-lang.ruhosting.nl/echo/ Dataset: https://hdl.handle.net/1839/00-0000-0000-0001-4892-C
Institution	Max Planck Insitute for Psycholinguistics, Radboud University Nijmegen, University of Lund

Figure 1: Example of a corpus entry in the compendium. Shown are the header, menu, free-form description and part of the info table. Not shown are the tables of used tasks and the list of references.

corpus-task pairs (Section 4.5). Language entries are described in Section 4.6.

All entries are interconnected, providing links between related resources, between languages and resources and between tasks and corpora. An example corpus entry can be seen in Figure 1. An example elicitation task entry including corpus-task pairs is shown in Figure 2.

4.1. Categories for Corpora

The following info table categories are provided for each corpus:

Languages: The languages used in the primary data of the corpus. Does not include languages used in annotation or translation.

Size: Size of the corpus. Depending on the information available, this may be specified as token count, type count, recording hours, number of video clips and/or file size.

Participants: Demographic information about the corpus participants. Apart from the number of participants this may include which regions they are from, age groups, gender distribution, and more. It is limited to demographic information that has been publicly documented.

Metadata Format: The file formats in which machine-readable metadata is provided by the corpus.

Translation: Which languages the primary data is translated into and how much of it has been translated.

Annotation: How much data has been annotated and which annotation conventions were used. If possible, a reference to the conventions is provided, otherwise information is paraphrased.

Data Format: The file formats in which the annotation/translation data of the corpus is provided.

4.2. Categories for Lexical Resources

The collection of lexical resources includes both lexical databases as well as electronic dictionaries. Lexical databases are language resources containing lexemes and additional information such as citation forms and translations. Dictionaries extend this information further, e. g. by documenting sign usage or sense disambiguation. As the boundaries between lexical database and dictionary are fluid, the compendium does not explicitly differentiate between the two.

Each lexical resource info table covers the following categories:

Languages: The languages used in the lexical resource. As most lexical resources can be used as bilingual dictionaries to some degree, this covers both signed and spoken languages.

Size: Number of lexical items. Items are identified as signs or types depending on the resource.

Linguistic Information: Which linguistic information is provided for lexical items, such as ID-glosses, translational equivalents, citation form video, meanings, phonetic transcription or categorisations, frequency and other statistics, list of corpus occurrences and more.

4.3. Categories for Corpora and Lexical Resources

The following info table categories apply to datasets in general, covering both corpora and lexical resources:

Licence: The licence conditions for using the dataset. These may be commonly used licences such as those by Creative Commons or custom licences defined for the dataset. A link to the licence is provided where possible.

Access: Describes how public and restricted data can be accessed. If the dataset has both public and restricted parts, this category identifies which parts of it are public.

Webpages: A list of relevant websites, such as those for the project, the research dataset, or portals for access by the general public.

Institutions: List of the universities or other organisations by which the dataset was created.

References: Important bibliographic references for the resource. If an external list of publications for the resource exist, a link to it is included here.

4.4. Categories for Data Collection Tasks

During corpus data collection, participants are guided by a series of tasks, such as retelling a story or open discussion of a given topic. The compendium lists data collection tasks used in multiple corpora. This information is intended to help with finding corpora that have comparable contents.

The info tables for data collection tasks cover the following categories:

Stimulus: Brief description of the stimulus provided to participants.

Target: The linguistic phenomena that the task is intended to elicit.

Degree of Interaction: An estimate whether the task usually results in a low, medium or high amount of interaction between participants. A reason for the degree may be given as a comment.

Duration: An estimate of how long the task usually lasts, based on instances observed in corpus data or published documentation.

Source: References to the material used in the task (e. g. books, films) or to scientific publications providing a definition of the task.

4.5. Categories for Corpus-Task Pairs

In addition to general descriptions of corpora and the data collection tasks that are used in their creation, the compendium also includes additional tables that provide information on the use of a task in a specific table. These tables contain the following categories:

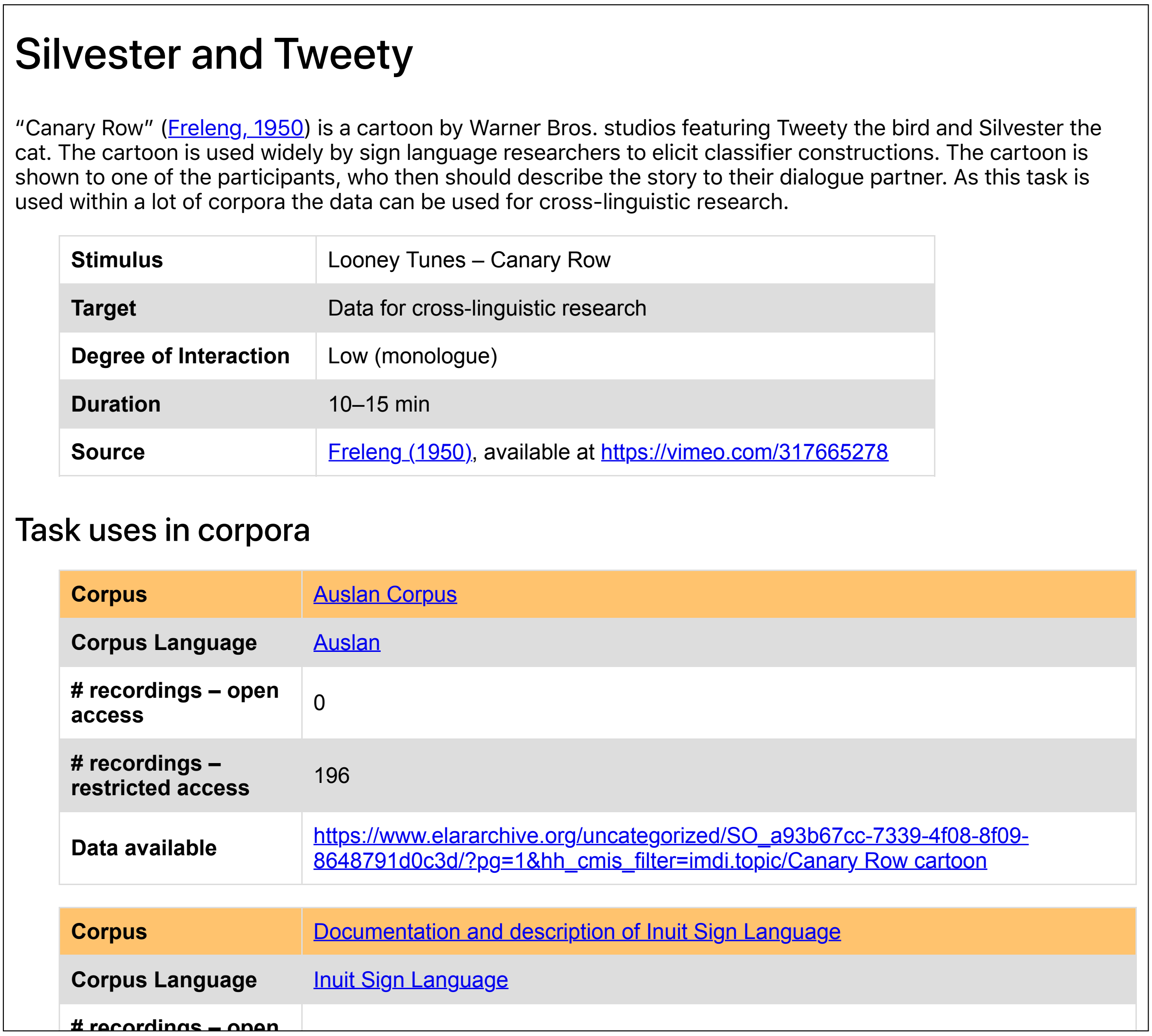
# recordings – open access: The number of recordings that are available in the publicly accessible part of the corpus.

# recordings – closed access: The number of recordings that are only available in the non-public part of the corpus.

Data available: Links to the corpus recordings of this task, where available. Where possible these links will connect only to the given task; otherwise disambiguating notes are provided to help find the task on the referenced page.

4.6. Languages

The compendium provides an index of the languages covered by its resources. As sign languages often go by a number of different names and acronyms, each language is given an entry that lists various common names and identifiers for it:



Silvester and Tweety

“Canary Row” (Freleng, 1950) is a cartoon by Warner Bros. studios featuring Tweety the bird and Silvester the cat. The cartoon is used widely by sign language researchers to elicit classifier constructions. The cartoon is shown to one of the participants, who then should describe the story to their dialogue partner. As this task is used within a lot of corpora the data can be used for cross-linguistic research.

Stimulus	Looney Tunes – Canary Row
Target	Data for cross-linguistic research
Degree of Interaction	Low (monologue)
Duration	10–15 min
Source	Freleng (1950), available at https://vimeo.com/317665278

Task uses in corpora

Corpus	Auslan Corpus
Corpus Language	Auslan
# recordings – open access	0
# recordings – restricted access	196
Data available	https://www.elararchive.org/uncategorized/SO_a93b67cc-7339-4f08-8f09-8648791d0c3d/?pg=1&hh_cmis_filter=imdi.topic/Canary Row cartoon

Corpus	Documentation and description of Inuit Sign Language
Corpus Language	Inuit Sign Language
# recordings – open	

Figure 2: Example of a data collection task entry in the compendium. Shown are the free-form description, info table and the first task use tables. Not shown are header, menu and list of references.

ISO 639-3: The unique identifier of the language in the ISO 639-3 code table.

Glottolog: The unique glottocode identifier of the language in the Glottolog database (Forkel and Hammarström, 2021).

Acronyms: Language acronyms commonly used by the language community or in research publications.

English names: English names for the language.

Local names: Names for the language used in its native region. So far this is limited to languages with a written form, which unfortunately prevents the representation of sign language names in their own language. For names written in other scripts than the latin alphabet, a transliteration is also provided.

Acronyms, English and local names are each sorted roughly by which variants are preferred within the language community and by how commonly they are used locally and in research. The most preferred English name and (where applicable) most preferred acronym of each language are shown in the language index. However, each language can still be found by all its other names, acronyms and identifiers by typing them in the provided search filter.

5. Conclusion

The *Sign Language Dataset Compendium* provides an extensive overview of corpora and lexical resources for many different signed languages from across the world. In addition it identifies a number of data collection tasks that have been used across different corpora. Information for each resource is presented in a standardised structure that is nevertheless flexible.

The compendium supports researchers in identifying language resources suitable to their needs, particularly in the case of cross-lingual research and research combining and comparing multiple resources. The compendium also highlights the imbalance in data availability across different languages while at the same time supporting the visibility of languages that are less often considered for data-driven sign language research.

The compendium is a growing resource that will be updated regularly. At the current time it contains 102 resources and 27 tasks, but more will be added over time. Various aspects of the compendium will be reevaluated as it grows, including the curation criteria and table categories described in this article. Possible improvements to the table categories that are under consideration are the addition of a recording setup category to describe factors like camera angles and the restructuring of the *linguistic information* category into more fine-grained categories.

The web format of the compendium will also receive additional feature updates. Plans for these include additional filter and sorting functions, e. g. for finding public datasets, and integrating machine-readable metadata standards.

Should you spot any inaccuracies, be able to contribute missing information or know of additional resources that should be included in the compendium, please contact us at sldc@dgs-korpus.de.

6. Acknowledgements

We would like to thank the following people for their valuable input, advice, and their information on indivdual resources: Ronice Müller de Quadros, Csilla Bartha, Eleni Efthimiou, Hope E. Morgan, Johanna Mesch, Katalin Liszi, and Thomas Troelsgård.

This work is supported in part by the EASIER (Intelligent Automatic Sign Language Translation) Project. EASIER has received funding from the European Union's Horizon 2020 research and innovation programme, grant agreement n° 101016982.

This work has in part been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities

7. Bibliographical References

Bird, S. and Simons, G. (2001). The OLAC metadata set and controlled vocabularies. In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources*, pages 7–18, Toulouse, France.

Calzolari, N., Soria, C., Del Gratta, R., Goggi, S., Quochi, V., Russo, I., Choukri, K., Mariani, J., and Piperidis, S. (2010). The LREC map of language resources and technologies. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 949–956, Valletta, Malta. European Language Resources Association (ELRA).

Forkel, R. and Hammarström, H. (2021). Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, Preprint. DOI: 10.3233/SW-212843.

Goosen, T. and Eckart, T. (2014). Virtual Language Observatory 3.0: What's new? In *Selected papers from the CLARIN 2014 Conference*, Soesterberg, Netherlands.

Hanke, T., Schulder, M., and Kopf, M. (2021). The sign-lang@LREC Anthology. Web Archive.

Hartzell, E. (2022). Egyptian minority language resources. DOI: 10.5281/zenodo.6385150.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131. DOI: 10.1075/ijcl.15.1.05joh.

Konrad, R. (2012). Sign language corpora survey. DOI: 10.25592/uhhfdm.9893.

Kopf, M., Schulder, M., and Hanke, T. (2021). Overview of datasets for the sign languages of Europe. Project Deliverable EASIER D6.1, EASIER Consortium. DOI: 10.25592/uhhfdm.9560.

Kopf, M., Schulder, M., Hanke, T., and Bigeard, S. (2022). Specification for the harmonization of sign language annotations. Project Deliverable EASIER D6.2, EASIER Consortium. DOI: 10.25592/uhhfdm.9842.

Moryossef, A. and Goldberg, Y. (2021). Sign language processing. Website.

Schmaling, C. H. (2012). Dictionaries of African sign languages: An overview. *Sign Language Studies*, 12(2):236–278. DOI: 10.1353/sls.2011.0025.

Simons, G. and Bird, S. (2003). The Open Language Archives Community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128. DOI: 10.1093/llc/18.2.117.

Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: the Virtual Language Observatory. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 1029–1034, Istanbul, Turkey. European Language Resources Association (ELRA).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018. DOI: 10.1038/sdata.2016.18.