Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

easier

# The Sign Language Dataset Compendium

## Creating an Overview of Digital Sign Language Resources

Maria Kopf, Marc Schulder, Thomas Hanke
University of Hamburg, Institute of German Sign Language and Communication of the Deaf, Germany

## Why?

- Identification of suitable sign language datasets challenging
- No single source of information
- Extensive literature review or word-of-mouth needed
- Amount of information varies widely
- Information distributed across different publications, data repositories and (potentially defunct) project websites

## What is Described?

### For Corpora

**Languages**: Languages of primary data; not including languages of annotation/ translation

**Size**: Token count, type count, recording hours, number of video clips, and/or file size

**Participants**: Publicly documented demographic information, such as number of participants, regions, age groups, gender distribution, and more

**Metadata Format**: File formats of machine-readable metadata

**Translation**: Amount and language of translations

**Annotation**: Amount of annotations, and conventions (paraphrased and/or with reference to published conventions)

**Data Format**: File formats of annotation/ translation data

### For Lexical Resources

**Languages**: Signed and spoken languages used in resource

**Size**: Number of lexical items (signs or types)

**Linguistic Information**: ID-glosses, translational equivalents, citation form video, meanings, phonetic transcription or categorisations, frequency and other statistics, list of corpus occurrences, and more

### For Corpora as well as Lexical Resources

**Licence**: Commonly used licence conditions such as Creative Commons or custom licences, link to the licence

**Access**: Identification of public and restricted parts, description of access

**Webpages**: Relevant websites such as project page, research dataset, portals for access by the general public, and more

**Institutions**: Universities and other organisations creating the dataset

**References**: Important bibliographic references, links to external lists of publications

## The Compendium

- Extensive **overview of linguistic resources** for sign languages around the globe
- Naturally used language by signers with L1 language proficiency
- Info tables with **structured information** in thematic categories
- Standardised but flexible format through free-form description
- Machine-readable **metadata**
- Pointers to data, project websites, literature references, etc.
- **Interconnected** entries, providing links between resources, tasks, and languages
- Available as website and static document
- Growing resource with **regular updates**

## Current Size

42
Corpora

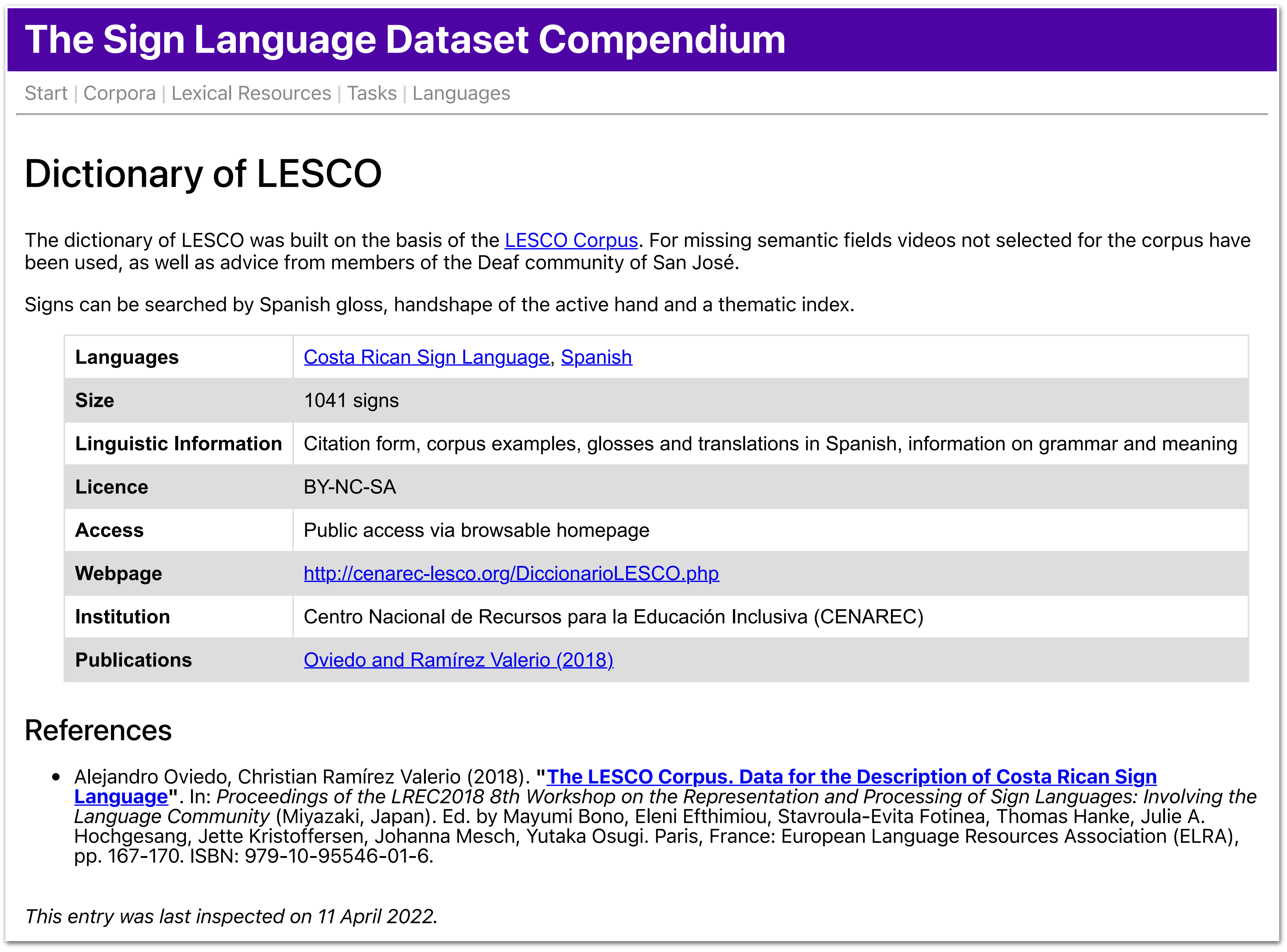
84
Lexical Resources

27
Collection Tasks

82
Sign Languages

## Where to Find It?

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



The Sign Language Dataset Compendium

Start | Corpora | Lexical Resources | Tasks | Languages

Dictionary of LESCO

The dictionary of LESCO was built on the basis of the LESCO Corpus. For missing semantic fields videos not selected for the corpus have been used, as well as advice from members of the Deaf community of San José.

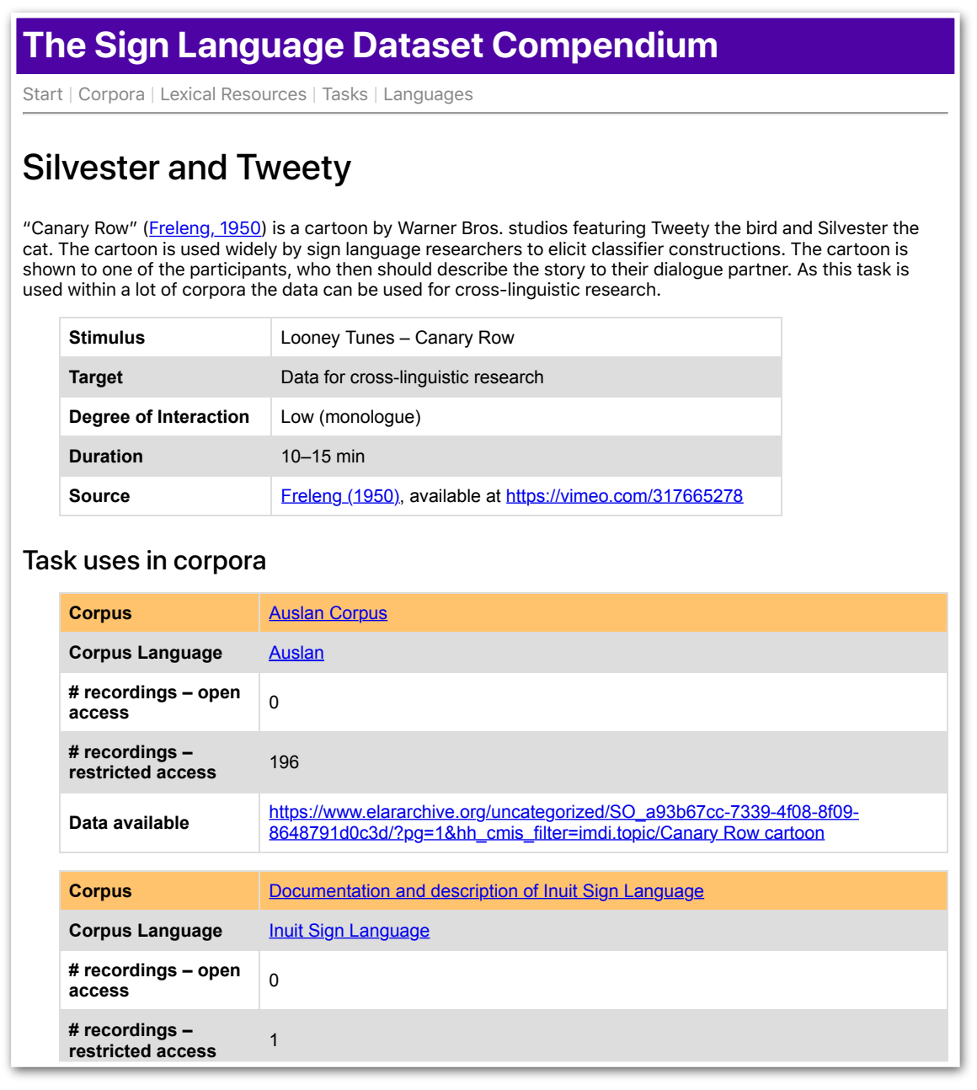
Signs can be searched by Spanish gloss, handshape of the active hand and a thematic index.

| | |
|---|---|
| Languages | Costa Rican Sign Language, Spanish |
| Size | 1041 signs |
| Linguistic Information | Citation form, corpus examples, glosses and translations in Spanish, information on grammar and meaning |
| Licence | BY-NC-SA |
| Access | Public access via browsable homepage |
| Webpage | http://cenarec-lesco.org/DiccionarioLESCO.php |
| Institution | Centro Nacional de Recursos para la Educación Inclusiva (CENAREC) |
| Publications | Oviedo and Ramírez Valerio (2018) |

References

- Alejandro Oviedo, Christian Ramírez Valerio (2018). "The LESCO Corpus. Data for the Description of Costa Rican Sign Language". In: *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community* (Miyazaki, Japan). Ed. by Mayumi Bono, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, Johanna Mesch, Yutaka Osugi. Paris, France: European Language Resources Association (ELRA), pp. 167-170. ISBN: 979-10-95546-01-6.

*This entry was last inspected on 11 April 2022.*



The Sign Language Dataset Compendium

Start | Corpora | Lexical Resources | Tasks | Languages

Silvester and Tweety

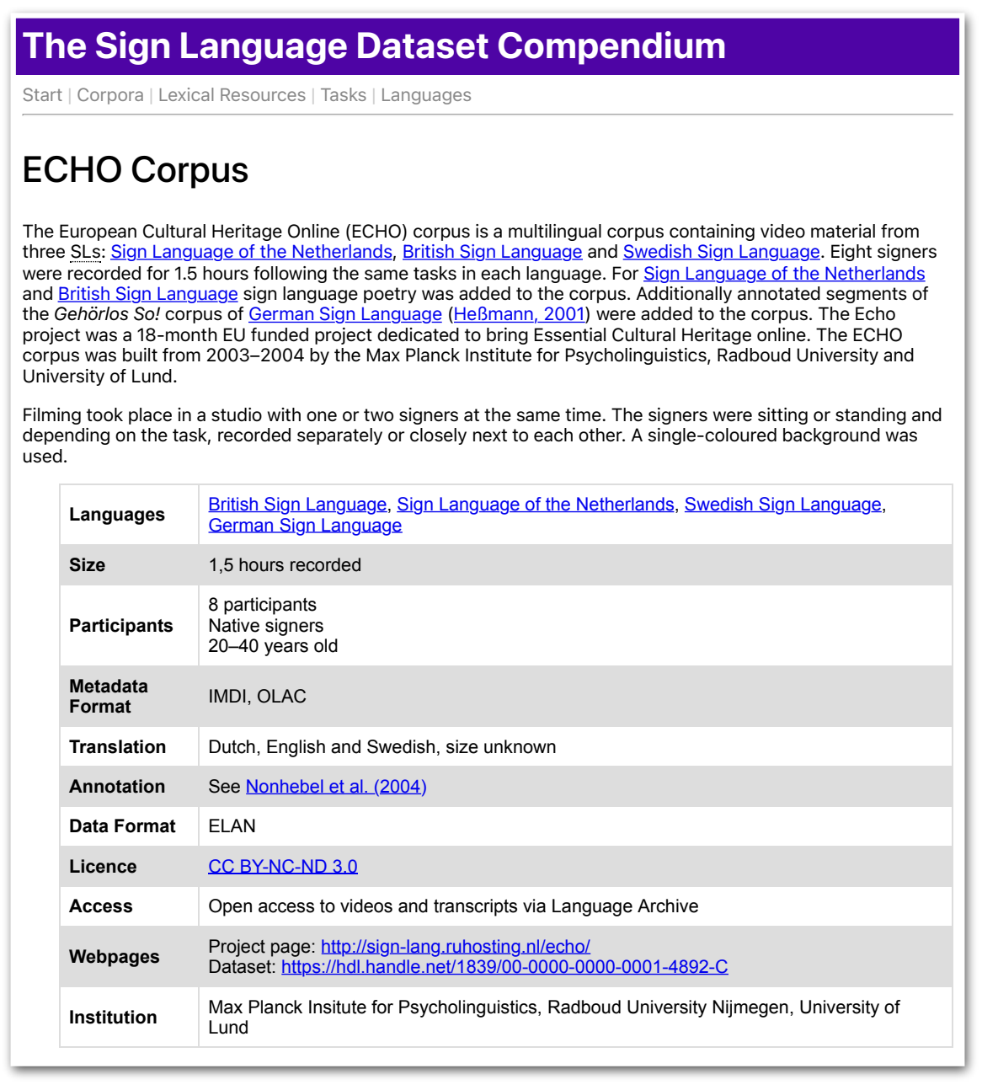
"Canary Row" (Freleng, 1950) is a cartoon by Warner Bros. studios featuring Tweety the bird and Silvester the cat. The cartoon is used widely by sign language researchers to elicit classifier constructions. The cartoon is shown to one of the participants, who then should describe the story to their dialogue 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 access | 0 |
| # recordings – restricted access | 1 |



The Sign Language Dataset Compendium

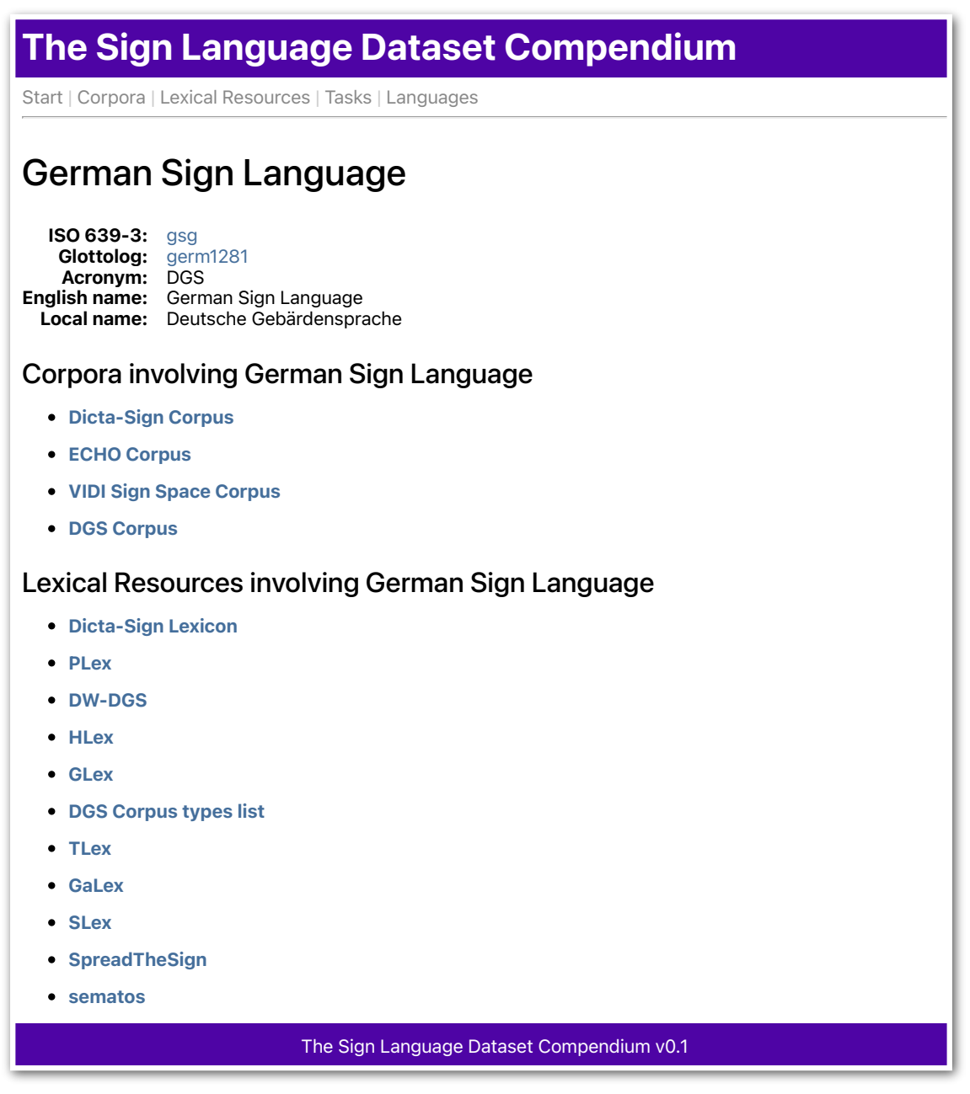
Start | Corpora | Lexical Resources | Tasks | Languages

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<br>Native signers<br>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/<br>Dataset: https://hdl.handle.net/1839/00-0000-0000-0001-4892-C |
| Institution | Max Planck Insitute for Psycholinguistics, Radboud University Nijmegen, University of Lund |



The Sign Language Dataset Compendium

Start | Corpora | Lexical Resources | Tasks | Languages

German Sign Language

| | |
|---|---|
| ISO 639-3: | gsg |
| Glottolog: | germ1281 |
| Acronym: | DGS |
| English name: | German Sign Language |
| Local name: | Deutsche Gebärdensprache |

Corpora involving German Sign Language

- Dicta-Sign Corpus
- ECHO Corpus
- VIDI Sign Space Corpus
- DGS Corpus

Lexical Resources involving German Sign Language

- Dicta-Sign Lexicon
- PLex
- DW-DGS
- HLex
- GLex
- DGS Corpus types list
- TLex
- GaLex
- SLex
- SpreadTheSign
- sematos

The Sign Language Dataset Compendium v0.1

## How?

### Method

- Literature review of 363 publications in the sign-lang@LREC Anthology
- Inspection of
  - Additional literature
  - Datasets
  - Project websites
- Personal correspondence with data creators
- Language specific curation criteria depending on size and number of available resources

### Curation Criteria

**General criteria:**

1. Must include video data
2. No sign-supported systems
3. No language acquisition data
4. No historical sign languages
5. Data must be attainable

**For corpora**

6. Must be (semi-)spontaneous signing
7. L1 signers
8. Data must have at least a partial translation and/or gloss annotation
9. At least 5 hours (minimum) or 10 hours (strict) of sign language recordings.

**For Lexical Resources:**

10. Must include index
11. At least 100 (minimum) or 1000 (strict) different signs.

**For data collection tasks:**

12. Used by multiple resources

### What are Commonly Used Tasks?

**Stimulus**: Description of stimulus provided to participants

**Target**: Linguistic phenomena intended to elicit

**Degree of Interaction**: Estimate of amount of interaction, reason for given degree

**Duration**: Estimate of duration of the task, based on instances observed in corpus data or published documentation

**Source**: References to material (e.g. books, films) or related scientific publications

### Which Corpus Uses which Task?

**# recordings – open access:** Number of recordings publicly available

**# recordings – closed access:** Number of recordings non-publicly available

**Data available:** Links to corpus recordings of this task with disambiguating notes to help find the task on the referenced page

Poster presented at the 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources. LREC 2022. Marseille, France. June 25th, 2022.
Updated for the EASIER Autumn School and Digital Total, Hamburg, September, 2023.

Funded by the Horizon 2020 Framework Programme of the European Union

We would like to thank the following people for their valuable input, advice, and their information on individual 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