Making Sign Language Resources Findable and Comparable: **The Sign Language Dataset Compendium**

Maria Kopf, Marc Schulder, Thomas Hanke
Institute of German Sign Language
Hamburg University

Background

- Sign languages are under-resourced
- Existing datasets hard to find
- Metadata & documentation even harder to find

- *First goal:* Overview of corpora & lexical resources for sign languages of Europe

- *New goal:* Global overview for all sign languages

The Sign Language Dataset Compendium

Start | About | Corpora | Lexical Resources | Tasks | Languages | Credit

The Sign Language Dataset Compendium

Welcome to the *Sign Language Dataset Compendium*, an overview of digital resources for signed languages suitable for research. The compendium covers both corpora and lexical resources. It also provides an overview of commonly used data collection tasks and in which corpora they were used. For those looking for datasets for a specific language, a language index is provided.

Should you know of additional resources, know of information that is missing from an entry, spot inaccuracies or wish to provide us any other feedback please contact us at sldc@dgs-korpus.de

About

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. As we follow the terminology of each individual resource differences in terminology, such as different size indication (sign, token, type) or the use of deaf vs. Deaf, may occur. Where possible we use consistent terminology, enriched with comments if needed. All entries are interconnected, providing links between related resources, between languages and resources and between tasks and corpora. Resources can be filtered using keywords.

How to Cite

To credit the compendium, please cite the following paper:

Kopf, M., Schulder, M., & Hanke, T. (2022). The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources. Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources, pp. 102–109.

BibTeX

```
@inproceedings{kopf:22025:sign-lang:lrec,
  author    = {Kopf, Maria and Schulder, Marc and Hanke, Thomas},
  title     = {The {Sign} {Language} {Dataset} {Compendium}: Creating an Overview of Digital Linguistic Resources},
  pages     = {102--109},
  editor    = {Efthimiou, Eleni and Fotinea, Stavroula-Evita and Hanke, Thomas and Hochgesang, Julie A. and Kristoffersen, Je
tte and Mesch, Johanna and Schulder, Marc},
  booktitle = {Proceedings of the {LREC2022} 10th Workshop on the Representation and Processing of Sign Languages: Multilingu
```

Language Documentation and Archiving Conference 2022

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Compendium Demo
Introduction

The Sign Language Dataset Compendium

Corpus

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

Language Documentation and Archiving Conference 2022

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Compendium Demo
Tasks

Common tasks used in this corpus

Task	Lexical elicitation
Corpus Language	British Sign Language
# recordings – open access	1
# recordings – restricted access	0
Data available	https://hdl.handle.net/1839/00-0000-0000-0001-49AF-B

Task	Lexical elicitation
Corpus Language	Sign Language of the Netherlands
# recordings – open access	4
# recordings – restricted access	0
Data available	https://hdl.handle.net/1839/00-0000-0000-0001-4A68-0

Task	Lexical elicitation
Corpus Language	Swedish Sign Language
# recordings – open access	1
# recordings – restricted access	0
Data available	https://hdl.handle.net/1839/00-0000-0000-0001-4AE2-C

Task	Retelling of fables
Corpus Language	British Sign Language
# recordings – open access	10
# recordings – restricted access	0
Data available	https://hdl.handle.net/1839/00-0000-0000-0001-4950-1

Task	Retelling of fables

Language Documentation and Archiving Conference 2022

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Compendium Demo
Tasks

The Sign Language Dataset Compendium

Start | About | Corpora | Lexical Resources | Tasks | Languages | Credit

Signed Languages | Spoken Languages

Languages

The datasets in the compendium involve 76 signed languages and 40 spoken languages.

Search: Find languages by their names, acronyms, ISO code, glottolog

Signed Languages

- **Adamorobe Sign Language (AdaSL)**: Adamorobe Sign Language Corpus Adamorobe Sign Language Lexicon
- **American Sign Language (ASL)**: Hallatlan Dictionary Dictio ASL Signbank SignStudy Langue Colorée ASL-LEX SpreadTheSign
- **Arabic Sign Languages (ArSLs)**: Langue Colorée SpreadTheSign
- **Argentine Sign Language (LSA)**: Señario de términos y expresiones básicas en la Lengua de Señas Argentina SpreadTheSign
- **Auslan**: Auslan Corpus Auslan Signbank SpreadTheSign
- **Australian Irish Sign Language (AISL)**: Australian Irish Sign Language
- **Austrian Sign Language (ÖGS)**: Dictio LedaSila SpreadTheSign
- **Belarusian Sign Language**: SpreadTheSign
- **Black American Sign Language (BASL)**: Black ASL Project Corpus
- **British Sign Language (BSL)**: Dicta-Sign Corpus ECHO Corpus British Sign Language Corpus Dicta-Sign Lexicon BSL SignBank SpreadTheSign
- **Bulgarian Sign Language**: SpreadTheSign
- **Cambodian Sign Language (CBDSL)**: Krousar Thmey Dictionary

Language Documentation and Archiving Conference 2022

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Compendium Demo Languages

Curation Criteria

- Qualitative & quantitative criteria
- Minimum criteria for SLs with almost no data
- Stricter criteria for SLs with more data
- ☞ Improve visibility of less resourced SLs

Standardisation

- Standardised structure for all entries
- Mostly freeform text
- Semantic markup via XML

- Work in progress:
  - machine-readable metadata
  - Registration with OLAC

Sustainability

- Growing resource
- Maintenance planned for 5 years
- Develop sustainable long term strategy
- Webpage and PDF version

Thank you!

Visit the Compendium:

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

Contact us:

sldc@dgs-korpus.de