



Corpus of German Sign Language (DGS)

The Project

- Runtime: 2009–2023 (15 years)
- Goals:
 - Reference corpus of DGS.
 - High quality public corpus.
 - Corpus-based dictionary DGS–German.

The Reference Corpus

- 560 hours** of natural conversations in DGS.
- Variety of tasks: discussions, life experiences, retellings, jokes...
- Sign per sign gloss annotations (work ongoing).
- Translation into German and English.

The Public Corpus

- 50 hour** subset of reference corpus.
- Representative sample of tasks and participants.
- High quality annotations & translations.
- Access via community or research portal.
- Regular updates (more content, corrections,...)

FAIR

Principles of Good Open Data

- Make (meta)data human- & machine-readable.
- Use unique persistent identifiers (e.g. DOIs).
- Clear means of access.
- Permissive open licence.
- Use established open file formats.
- Strong metadata & documentation.



CARE

Principles of Ethical Open Data for Minorities

- Minority population should **benefit from data** and be **empowered to exert control** over it.
- Researchers responsible for public **accountability**.
- Rights and wellbeing primary concern at **all stages of data life cycle**.

Designed for Indigenous Peoples, applies to deaf communities, too.

CARE in the DGS Corpus

Data Collection

- Deaf team members & community focus group.
- Corpus designed as record of deaf culture.
- Informed consent in DGS & German.
 - Multiple options for data sharing conditions.
 - Participants could review recordings.
 - Permanent control over data publication.

Size of Public Corpus

- Public corpus data must be high quality.
- Avoid harm from bad translations/annotations.
- Anonymisation relies on good data.
- 1h of corpus = 1000h of work
- 50h good trade-off for representative sample.

Access & Licensing

- Easy public access.
- Protective licences to prevent harm:
 - My DGS*: Private use, teaching.
 - My DGS – annotated*: Only linguistic research.
 - Custom licence for other uses/full corpus access.

Community Contact

- Present progress at deaf community events.
- Regular focus group meeting.
- React to community feedback.
- DGS-centric social media presence.

Privacy

- Limit public information of participants: Region, age group, gender.
- Fully anonymise 3rd party information: Names, dates of birth, precise places of residence.
- Retain information internally to track authorship.

Different Formats for Different Needs

My DGS

Community Portal



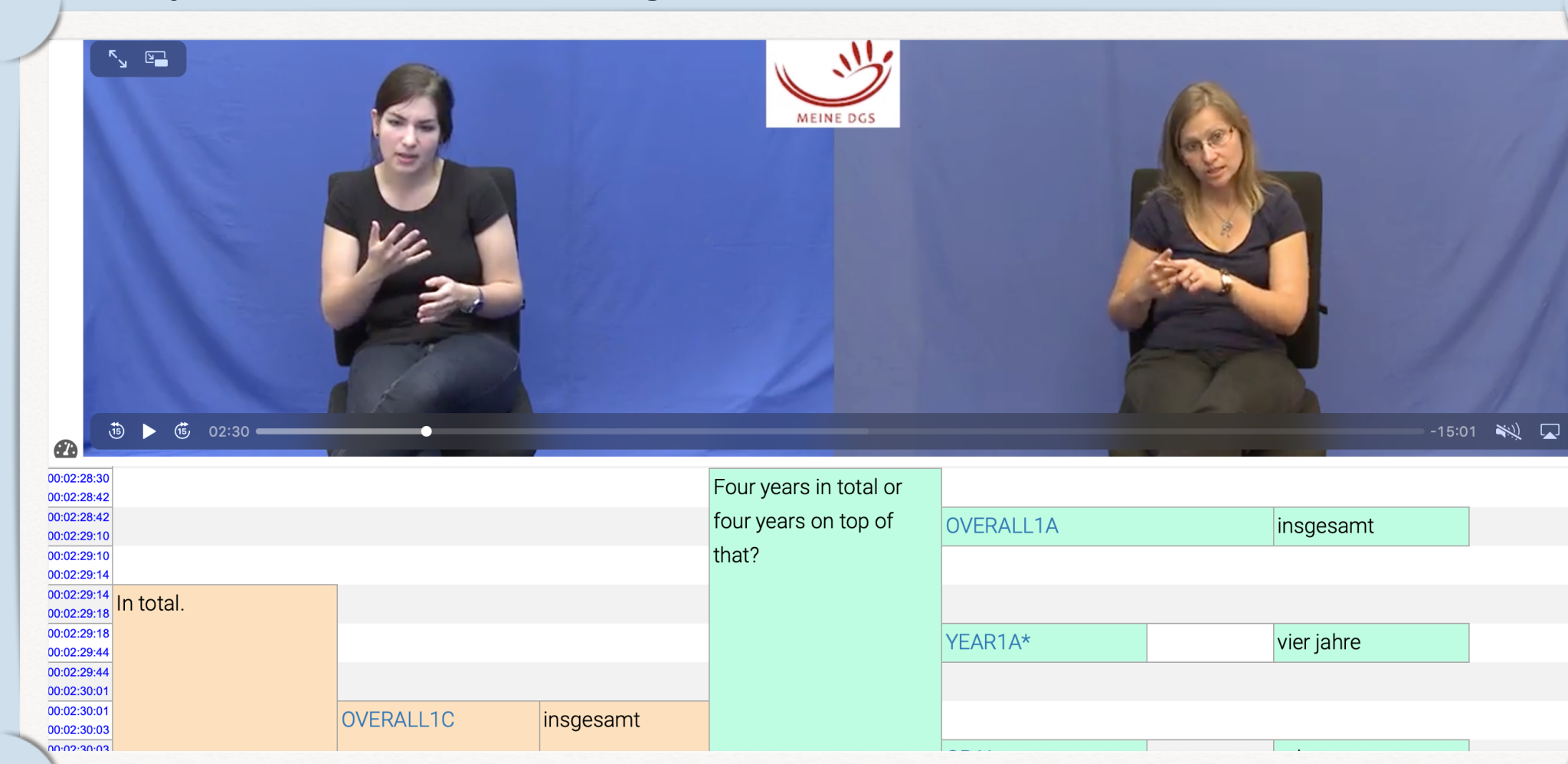
- Focus on finding interesting stories.
- Minimal text interface.
- Optional German subtitles.

My DGS – annotated

Research Portal



- Data and interface in English and German.
- Full annotations, translation, metadata, OpenPose.
- Type index of all signs.



FAIR in the DGS Corpus

Digital Object Identifiers (DOIs)

Unique persistent identifiers

Public Corpus DOIs:

- Each portal.
- Each transcript.
- Each sign type.
- Update releases get new DOIs.

More DOIs:

- Original recordings.
- Project notes, posters, presentations, code.

DOI Metadata

General Information

Öffentliches DGS-Korpus Release 3.0 / Public DGS Corpus Release 3.0

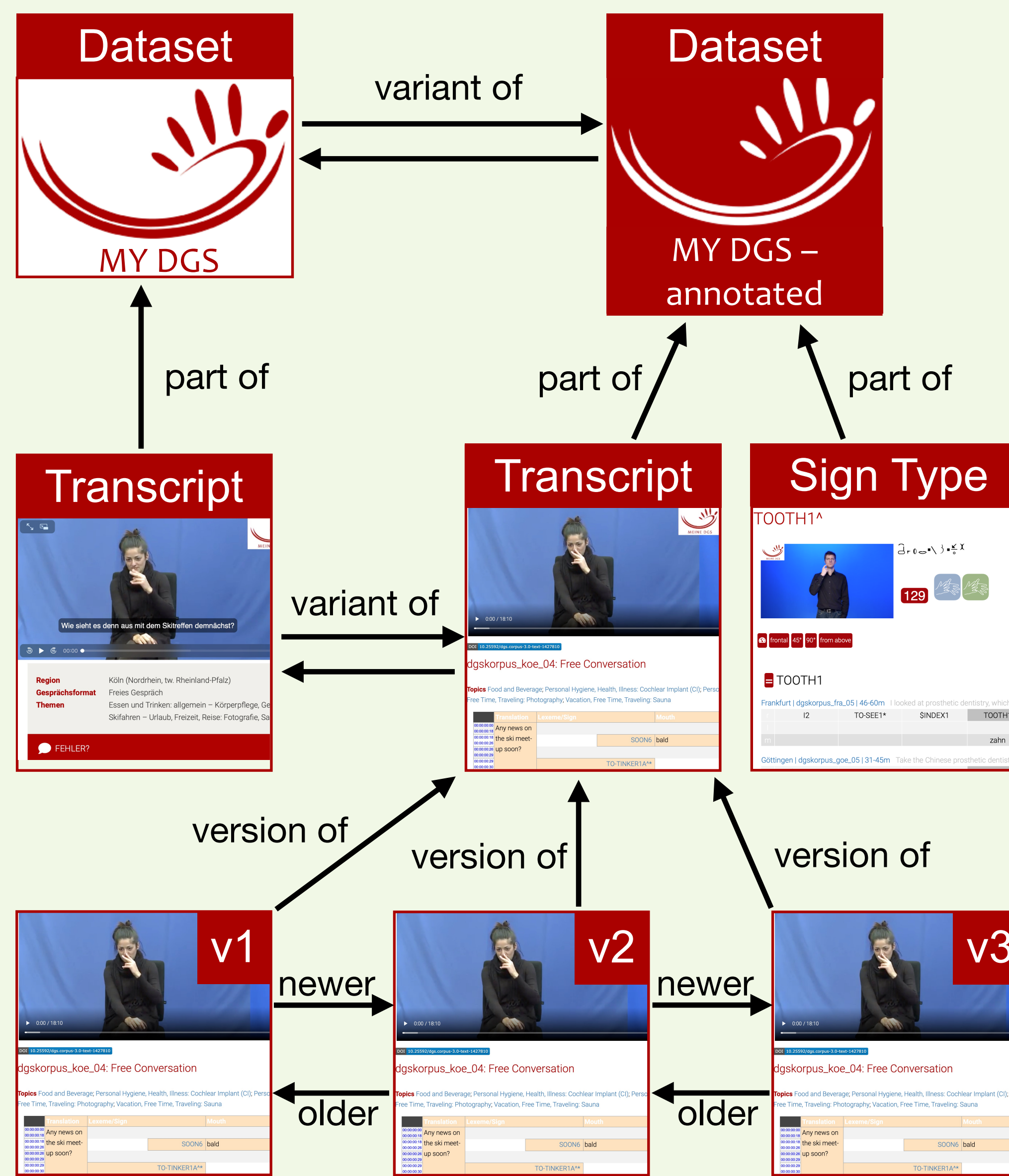
10.25592/dgs.corpus-3.0	
Dieses Release / This release:	
Öffentliches DGS-Korpus Release 3.0	Webseiten DE
Public DGS Corpus Release 3.0	Website EN
Ältere Versionen / Older versions:	
Öffentliches DGS-Korpus Release 2.0	10.25592/dgs.corpus-2.0
Öffentliches DGS-Korpus Release 1.0	10.25592/dgs.corpus-1.0
Public DGS Corpus Release 1.0	
Versionsunabhängige DOI / Version-independent DOI:	
Jeweiliges neuestes Release des Öffentlichen DGS-Korpus	10.25592/dgs.corpus
Always the newest release of the Public DGS Corpus:	
Zitiervorschlag / Cite as:	
König, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, T., Jezioraki, O., König, R., Nishio, R., Regen, A., Salden, U., Wagner, S., Woreck, S., Böse, O., Jahn, E.,	

dgskorpus_koe_05 – Erlebnisbericht / Experience Report

10.25592/dgs.corpus-3.0-text:1428038	
Diese Version / This version:	
Im Öffentlichen DGS-Korpus Release 3.0:	Webseite DE
In the Public DGS Corpus Release 3.0:	Web Page EN
ILex	1428038_ilex
ELAN	1428038_elan
Video A1	1428038_A1.mp4
Video B1	1428038_B1.mp4
Video C	1428038_C.mp4
SRT	1428038_srt
OpenPose	1428038_opense
Metadata (CMDI)	

Named Relations

Interconnected DOIs



Documentation

- Peer-reviewed publications.
- 30+ project notes,
 - includes *data statement*.
- CMDI metadata for each transcript.

Archival

- Long term storage in university data repository.
- Provides redundant storage & DOIs.
- Archival of original recordings & documentation.

CMDI Metadata

Language Data Information

