# CARE, FAIR and the DGS Corpus

Implementing Ethical Open Data Practices in a Large Sign Language Dataset

**Marc Schulder & Thomas Hanke**
Institute of German Sign Language
Universität Hamburg

# CARE, FAIR and the DGS Corpus

## Implementing Ethical Open Data Practices in a Large Sign Language Dataset

DGS

=

**D**eutsche **G**ebärden**s**prache

=

German Sign Language

# The DGS Corpus

- DGS-Korpus Project: 2009-2023
- Reference corpus: 560 hours
- Public corpus: 50 hours

- Sign for sign annotation
- English & German translations
- 1h recording = 800h–1000h work

# The DGS Corpus

- Stakeholders: Deaf community in Germany
- “Nothing about us without us”
- Deaf project members & focus group

- Linguistic research
- Record of deaf culture

# The DGS Corpus

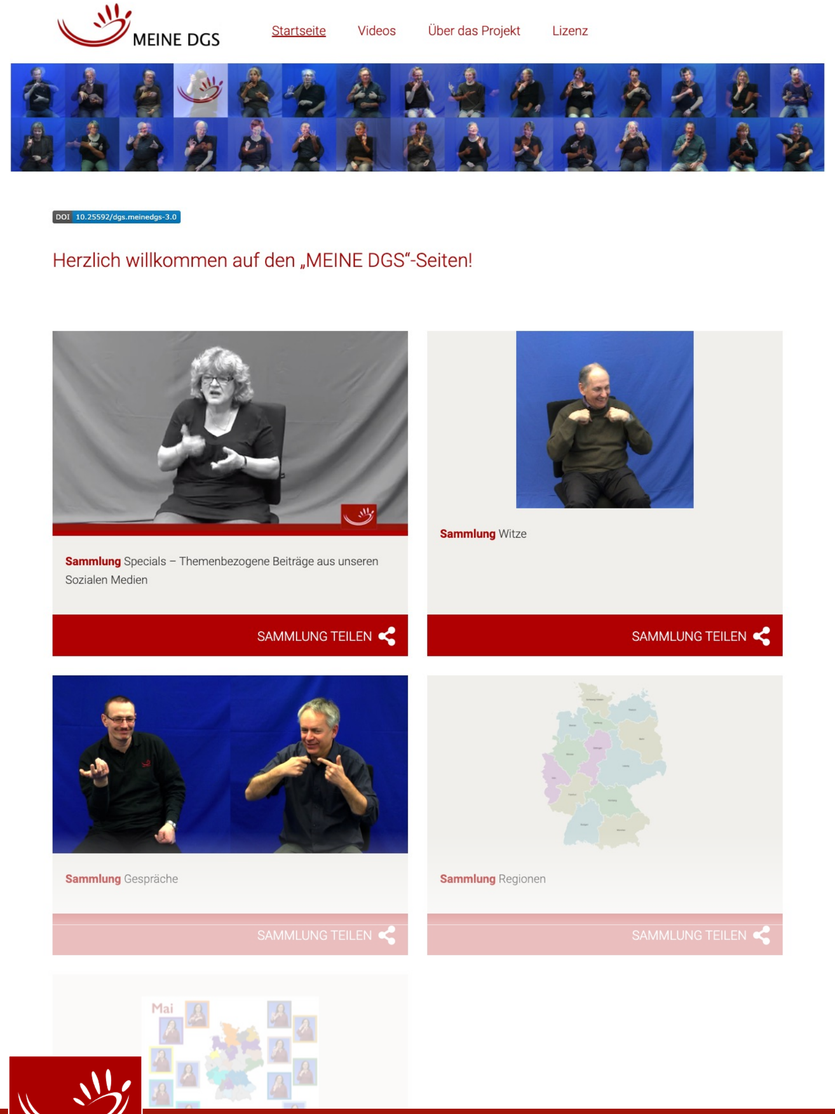
- 330 participants from all across Germany
- Informed consent (in DGS & German)
  - Project goals
  - Data uses
  - Participant rights
    - Restricting access
    - Approve content

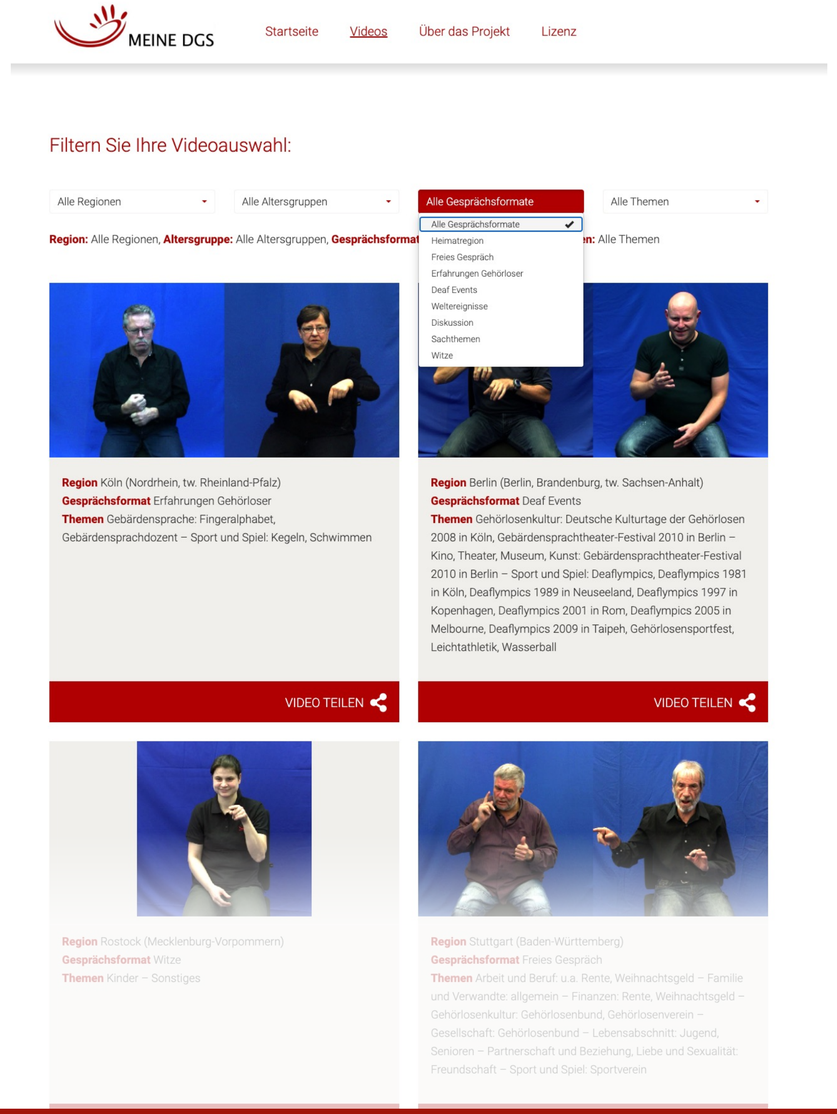
# The DGS Corpus

- First full release: 2018
- Update releases: 2019, 2020, 2022

- Different portals for different needs:
  - Community Portal
  - Research Portal

# My DGS
## Community Portal



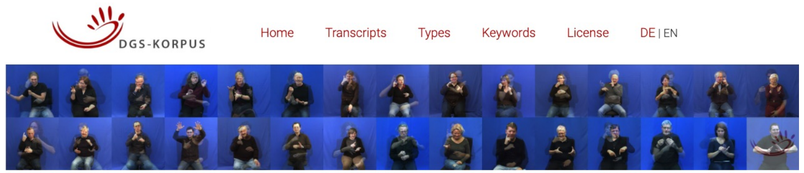
# Filtern Sie Ihre Videoauswahl:

Nürnberg | Alle Altersgruppen | Erfahrungen Gehörloser | Alle Themen

**Region:** Nürnberg, **Altersgruppe:** Alle Altersgruppen, **Gesprächsformate:** Erfahrungen Gehörloser, **Themen:** Alle Themen

| | |
|---|---|
| **Region** | Nürnberg (Bayern-Nord) |
| **Gesprächsformat** | Erfahrungen Gehörloser |
| **Themen** | Gebärdensprache: Internationale Gebärdensprache – Gehörlosenkultur: Identität als Gehörloser – Gesellschaft: Identität als Gehörloser – Kommunikation: Kommunikation mit Hörenden – Lebensabschnitt: Schulzeit – Schule und Unterricht: Schulzeit – Urlaub, Freizeit, Reise: Neuseeland |

# My DGS – annotated
## Research Portal




DOI 10.25592/dgs.corpus-3.0 on the way to release 4

### Welcome to the Public DGS Corpus Release 3!

In this portal you find 50 hours of video materials from the DGS-Korpus project made available together with annotations for research purposes.

If you want to download materials, please pay attention to the license conditions.

By clicking Transcripts, you can view the available data sorted by either transcript name or elicitation format. Several download links are available (cf. below, File Formats available for Download). By clicking the transcript name, you can open an online preview of the transcript. If you want to browse the videos first without paying attention to the annotation and you can read German, you find them in the MEINE DGS portal where they are available with subtitles.

By clicking Types, you can view the list of all sign types occurring in the public corpus. Click on one of these types to see all corresponding tokens in the public corpus. Clicking once again on a token reference brings directly to the occurrence in the transcript.

For all transcripts, keywords have been assigned in order to provide a rough content-related access to the data. By clicking Keywords, you can view an index of all keywords and find the transcripts they have been assigned to.

### Background Information

### Data Elicitation Formats

We used a set of 20 different tasks for the informants. The formats ranged from story-retelling (with prompts in sign, picture, or movie) to discussions on a given topic as well as free conversations. With careful planning, it was possible to make the mix of formats diverting enough that most participants enjoyed the recording session despite a net length of 5 hours.

The set contains a number of tasks previously used in other corpus projects, both on spoken and sign languages, to lay a basis for cross-linguistic research, as well as new formats. Not all details of the newly developed elicitation materials are available in the publications in order to keep the material suitable for future data collections. The materials are, however, available to other researchers upon request.

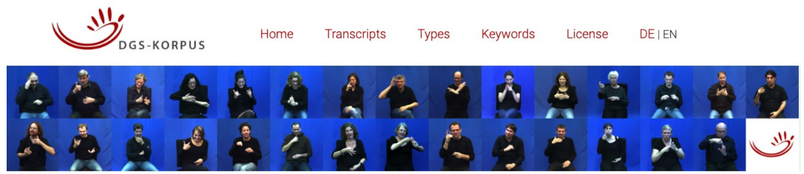
For more detail, please consult the following publications:

Hanke, Thomas / Hong, Sung-Eun / König, Susanne / Langer, Gabriele / Nishio, Rie / Rathmann, Christian (2010): "Designing Elicitation Stimuli and Tasks for the DGS Corpus Project". Poster presented at the Theoretical Issues in Sign Language Research Conference (TISLR 10), Sept 30 – Oct 2, 2010 at Purdue University, Indiana, USA. [Poster]
Nishio, Rie / Hong, Sung-Eun / König, Susanne / Konrad, Reiner / Langer, Gabriele / Hanke, Thomas / Rathmann, Christian (2010). "Elicitation methods in the DGS (German Sign Language) Corpus Project". Poster presented at the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, following the 2010 LREC Conference in Malta, May 22-23, 2010. Paris: ELRA, pp. 178-185. [Paper] [Poster]

### Data Collection Regions

From experiences in earlier projects, it was one of the key decisions to have a mobile studio to be set up in different places across Germany. The idea was to have as much of a "local" spirit with the recording location in the region and all persons involved coming from that region while still ensuring high-quality recordings needed for transcription. Obviously, the number of locations selected for recordings needs to be a compromise between, on the one hand, localness in the above sense and the informants' travel times, and, on the other hand, the logistics.

Our solution was the definition of thirteen data collection regions, trying to respect the catchment areas of current and former Schools for the [illegible]land) borders determining a. o. educational settings, especially the former border between West and East Germany, suspected [illegible]ut also practical considerations such as travel times to the recording locations. The regions were further subdivided into up to [illegible]evant for informant selection. Large metropolitan areas form their own sub-regions, in contrast to others with mixed or more

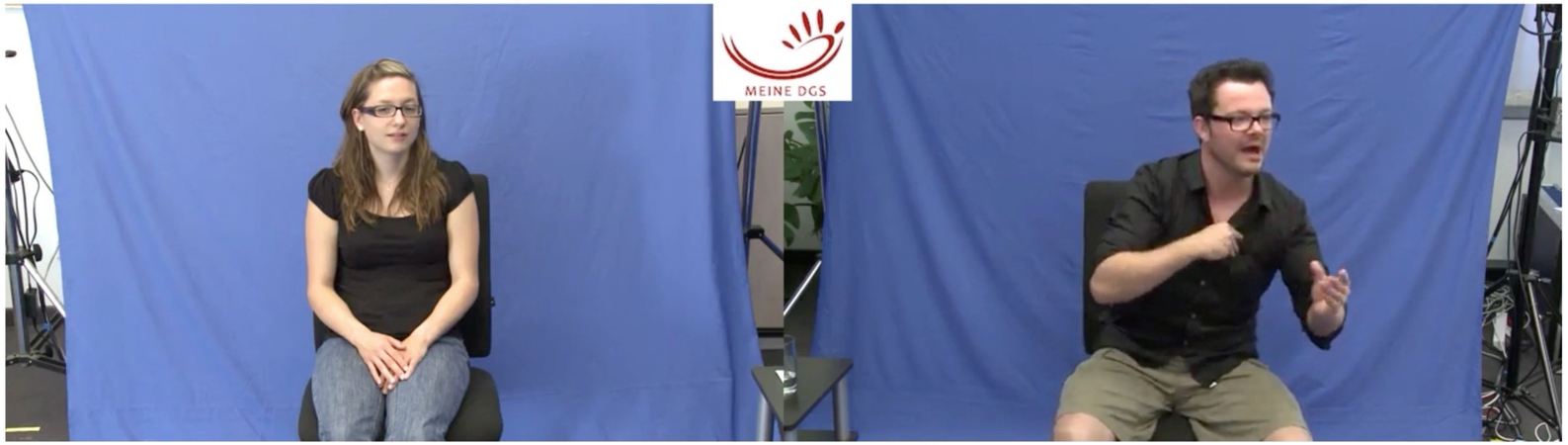



### Public Corpus Data

| ript | Age Group | Format | Topics | iLex | ELAN | Video A | Video B | Video Total | SRT | Video AB | CMDI | Op |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| us_ber_01 | 18-30 | Experience of Deaf Individuals | Communication: Communication with Hearing People<br>Deaf Culture<br>Family and Relatives: Family<br>Period of Life: Kindergarten<br>Period of Life: Schooldays<br>School and Education: Boarding School<br>School and Education: Schooldays<br>Society: Role Models<br>Work and Profession: Kindergarten | | | | | | | | | |
| us_ber_01 | 18-30 | Joke | Food (Joke) | | | | | | | | | |
| us_ber_01 | 18-30 | Experience of Deaf Individuals | Communication: Communication with Hearing People<br>Family and Relatives<br>Family and Relatives: Family<br>Personal Hygiene, Health, Illness: Cochlear Implant (CI)<br>Personal Hygiene, Health, Illness: Illness<br>Society: Hearing-Deaf<br>Vacation, Free Time, Traveling: Free Time | | | | | | | | | |
| us_ber_01 | 18-30 | Experience Report | Society: British Royal Family<br>Society: Lady Diana<br>Traffic: Accident<br>Traffic: Lady Diana | | | | | | | | | |
| us_ber_02 | 46-60 | Experience of Deaf Individuals | City and Country: Berlin<br>GDR: Escape<br>GDR: Fall of the Berlin Wall (1989)<br>Period of Life: Schooldays<br>Politics: Fall of the Berlin Wall (1989)<br>Relationship, Love and Sexuality: Meeting one's Partner<br>School and Education: School for Hard of Hearing<br>School and Education: School for the Deaf<br>School and Education: Schooldays<br>Society: Fall of the Berlin Wall (1989)<br>Sports and Games: Soccer<br>Sports and Games: Sports<br>Sports and Games: Sports Club<br>Work and Profession: Apprenticeship | | | | | | | | | |
| us_ber_02 | 46-60 | Free Conversation | Communication<br>Communication: Communication with Hearing People | | | | | | | | | |

DGS-KORPUS

Home Transcripts Types Keywords License DE | EN




| | | | |
|---|---|---|---|
| 00:05:01:08 | | | |
| 00:05:01:08 – 00:05:01:23 | person doesn't understand a sign, I can spell it, and the other one can show me the sign. | TO-MEET2B* | |
| 00:05:01:23 – 00:05:01:31 | | | |
| 00:05:01:31 – 00:05:01:34 | | $GEST^ | |
| 00:05:01:34 – 00:05:01:45 | | | |
| 00:05:01:45 – 00:05:01:46 | | I1 | |
| 00:05:01:46 – 00:05:02:04 | | | |
| 00:05:02:04 – 00:05:02:15 | | TO-SIGN1A | [MG] |
| 00:05:02:15 – 00:05:02:19 | | | |
| 00:05:02:19 – 00:05:02:21 | | WHAT1B* | [MG] |
| 00:05:02:21 – 00:05:02:30 | | | |
| 00:05:02:30 – 00:05:02:31 | | I1 | |
| 00:05:02:31 – 00:05:02:37 | | | |
| 00:05:02:37 – 00:05:02:49 | | ALPHABET1 | |
| 00:05:02:49 – 00:05:03:04 | | | |
| 00:05:03:04 – 00:05:03:15 | | $INDEX1* | [MG] |
| 00:05:03:15 | | | |
| | | TO-TELL3A* | [MG] |

# My DGS – annotated
## Sign Type Index

A-LEVEL1 (20 tokens) → SPIKY2A^

A-MOMENT-AGO1^ (186 tokens)

A-MOMENT-AGO1 (65 tokens) → A-MOMENT-AGO1^

A-MOMENT-AGO2 (42 tokens) → PAST1^

A-WHOLE-LOT1^ (134 tokens)

A-WHOLE-LOT2 (101 tokens) → A-WHOLE-LOT1^

AACHEN2 (1 token) → $ALPHA^

AACHEN3^ (13 tokens)

AACHEN3 (13 tokens) → AACHEN3^

ABDOMEN1 (1 token) → AREA1A^

ABM1^ (10 tokens)

ABM1 (10 tokens) → ABM1^

# My DGS – annotated
## Sign Type Index

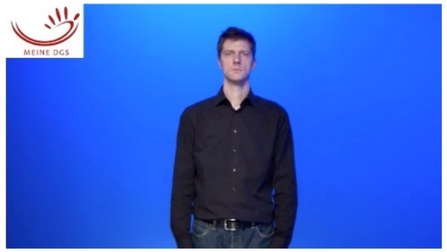
DOI 10.25592/dgs.corpus-3.0-type-18809

Be FAIR and CARE

Findable Accessible Interoperable Reusable

Collective Benefit Authority to Control Responsibility Ethics

TOOTH1^



129

frontal 45° 90° from above

TOOTH1

Frankfurt | dgskorpus_fra_05 | 46-60m I looked at prosthetic dentistry, which paid really well.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| r | I2 | TO-SEE1* | $INDEX1 | TOOTH1 | TECHNOLOGY1* | TOOTH2 | TECHNOLOGY1* |
| l | | | | | | | |
| m | | | | zahn | | zahntechnik | |

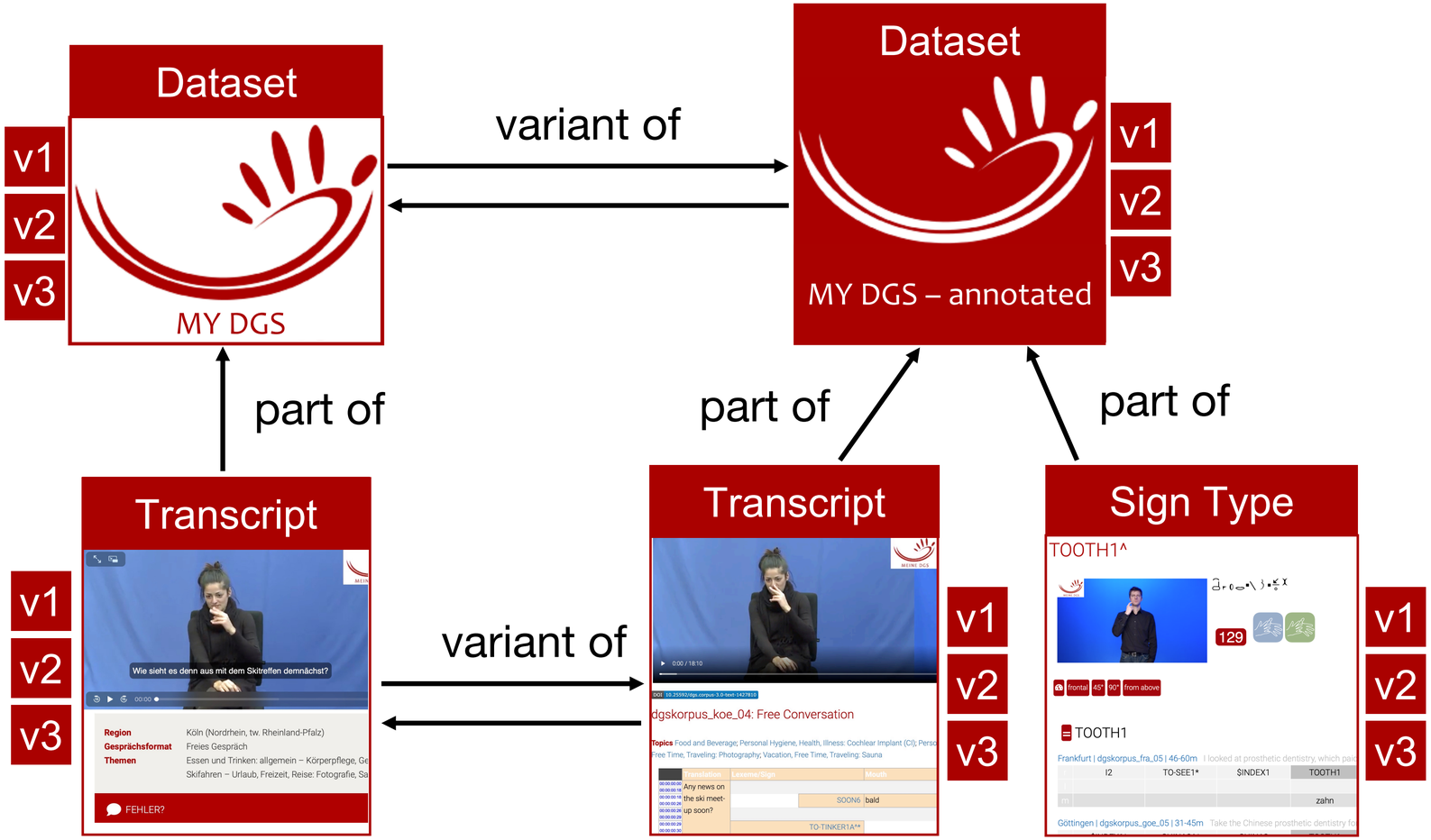
Göttingen | dgskorpus_goe_05 | 31-45m Take the Chinese prosthetic dentistry for instance.

| | | | | | |
|---|---|---|---|---|---|
| r | $INDEX1* | CHINA3A* | CHINA2 | TOOTH1 | TECHNOLOGY1 |
| l | | | | | |
| m | | | china | zahntechnik | |

Göttingen | dgskorpus_goe_05 | 31-45m They sell each other teeth.

| | | | | | |
|---|---|---|---|---|---|
| r | | | TO-SELL1* | TOOTH1 | TO-SPIN8^ |
| l | | | | | |
| m | | | verkaufen | zahn | |

# Persistent Identifiers



# Metadata

## DOI

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

doi 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: Public DGS Corpus Release 2.0: | 10.25592/dgs.corpus-2.0 |
| Öffentliches DGS-Korpus Release 1.0: Public DGS Corpus Release 1.0: | 10.25592/dgs.corpus-1.0 |

### Versionsunabhängige DOI / Version-independent DOI:

| | |
|---|---|
| Jeweils neuestes Release des Öffentlichen DGS-Korpus: Always the newest release of the Public DGS Corpus: | 10.25592/dgs.corpus |

### Zitiervorschlag / Cite as:

Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., Worseck, S., Böse, O., Jahn, E., Schulder, M. 2020. *MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release* [Dataset]. Universität Hamburg. https://doi.org/10.25592/dgs.corpus-3.0

## CMDI


```
▼<cmdp:Content>
  <cmdp:Genre>Discourse</cmdp:Genre>
  <cmdp:Task>Experience Of Deaf Individuals</cmdp:Task>
  <cmdp:Modalities>gestures; pointing-gestures; signs; eye-
  gaze; facial-expressions</cmdp:Modalities>
 ▼<cmdp:CommunicationContext>
    <cmdp:Interactivity>interactive</cmdp:Interactivity>
    <cmdp:PlanningType>spontaneous</cmdp:PlanningType>
    <cmdp:Involvement>non-elicited</cmdp:Involvement>
    <cmdp:SocialContext>Controlled
    environment</cmdp:SocialContext>
    <cmdp:EventStructure>Dialogue</cmdp:EventStructure>
    <cmdp:Channel>Face to Face</cmdp:Channel>
  </cmdp:CommunicationContext>
 ▼<cmdp:Content_Languages>
   ▼<cmdp:Content_Language>
      <cmdp:Id>ISO639-3:gsg</cmdp:Id>
      <cmdp:Name>Deutsche Gebärdensprache</cmdp:Name>
      <cmdp:Dominant>true</cmdp:Dominant>
      <cmdp:SourceLanguage>Unspecified</cmdp:SourceLanguage>
      <cmdp:TargetLanguage>Unspecified</cmdp:TargetLanguage>
    </cmdp:Content_Language>
  </cmdp:Content_Languages>
</cmdp:Content>
▼<cmdp:Actors>
 ▼<cmdp:Actor>
    <cmdp:Role>speaker/signer</cmdp:Role>
    <cmdp:Name>BER-36 (informant A)</cmdp:Name>
    <cmdp:Code>BER-36</cmdp:Code>
    <cmdp:FamilySocialRole>Unspecified</cmdp:FamilySocialRole>
    <cmdp:EthnicGroup>Unspecified</cmdp:EthnicGroup>
    <cmdp:BirthDate>Unspecified</cmdp:BirthDate>
    <cmdp:Sex>Female</cmdp:Sex>
    <cmdp:Education>Unspecified</cmdp:Education>
    <cmdp:Anonymized>true</cmdp:Anonymized>
   ▼<cmdp:Age>
     ▼<cmdp:AgeRange>
       ▼<cmdp:MinimumAge>
          <cmdp:years>18</cmdp:years>
        </cmdp:MinimumAge>
       ▼<cmdp:MaximumAge>
          <cmdp:years>30</cmdp:years>
```

# Documentation

## AP02 – Datenerhebung / Data Collection

AP02-2009-01 (de) Handbuch für Kontaktpersonen Teil I: Projekt, Werbung, Informantensuche, Raumsuche
AP02-2009-02 (de) Handbuch für Kontaktpersonen Teil II: Erhebung, Einverständniserklärung
AP02-2009-03 (de) Handbuch für Kontaktpersonen Teil III: SessionDirector
AP02-2010-01 (de) Erstellung der Elizitationsmaterialien: Anpassung an Lebenssituation in der ehemaligen DDR
AP02-2010-02 (de) Erstellung der Elizitationsmaterialien: Sachthemen
AP02-2010-03 (de) Ablauf Postproduktion
AP02-2010-04 (de) Umstellung auf HD-Stereokameras
AP02-2011-01 (de) Auswertung und Anpassung der Elizitationsmaterialien: Sachthemen
AP02-2012-01 (de) Auswertung und Anpassung der Elizitationsmaterialien: Sachthemen
AP02-2014-01 (en) H.264 and iLex

## AP03 – Transkriptionsmethodik / Transcription Methodology

AP03-2010-01 (de+en) Segmentierung / Segmentation
AP03-2018-01 (de+en) Öffentliches DGS-Korpus: Annotationskonventionen / Public DGS Corpus: Annotation Conventions

## AP04 – Tools

AP04-2011-01 (en) SessionDirector
AP04-2013-01 (en) Data Visualisation in iLex
AP04-2015-01 (en) Documentation of the Feedback-System and its Integration into iLex
AP04-2016-01 (en) Using d3js to Visualise Data in iLex
AP04-2018-01 (en) Processing DGS-Korpus Data with OpenPose on the Hamburg High Performance Cluster
AP04-2021-02 (de+en) HamNoSys in TeX

## AP05 – Basistranskription / Basic Transcription

AP05-2014-01 (de) Übersetzung
AP05-2017-01 (de) Formale und inhaltliche Prüfschritte zur Korpusveröffentlichung

## AP06 – Korpusveröffentlichung / Corpus Publishing

AP06-2013-01 (de) Auswahl von Aufnahmen für das Teilkorpus
AP06-2019-01 (en) OpenPose in the Public DGS Corpus
AP06-2020-01 (en) Data Statement for the Public DGS Corpus

## AP07 – Grundwortschatz / Basic Vocabulary

AP07-2014-01 (de) Entwicklung des DGS Feedback-Systems

## AP10 – Wörterbuchinhalte / Dictionary

# Archival

- Original recordings: 730,000 GB
- Supplementary data: 250,000 GB

- Archival in UHH Research Data Repository:
  - Invenio RDM
  - Persistent storage
  - Public metadata

# Thank You!

- Check out the corpus:
  - My DGS
    http://meine-dgs.de
  - My DGS — annotated
    http://ling.meine-dgs.de