Publishing DGS Corpus Data:
Different Formats for Different Needs

Elena Jahn, Reiner Konrad, Gabriele Langer, Sven Wagner, Thomas Hanke
University of Hamburg, Institute of German Sign Language and Communication of the Deaf

DGS-KORPUS

UH Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

AKADEMIE DER WISSENSCHAFTEN IN HAMBURG

The Public DGS-Corpus
- About 50 hrs of signed conversations selected out of 560 hrs from the DGS-Korpus data
- Pairs of informants presented side by side
  - Studio: informants facing each other with moderator in between
- Informants: 327 out of 330
- **Free online access via three different formats**

Selection Criteria
- Balanced for region, sex, and age
- Nearly all elicitation tasks (18 out of 20)
- Large variety of topics
- Different styles of signing

Selection Process
- Rating of elicitation tasks according to their importance for Deaf community
- ➡ Prioritised list of 7 tasks (47 hrs):
  - Experiences as a Deaf person, Free conversation, Discussion, Subject areas, Experience reports, Region of origin, Deaf events.
  - Plus Jokes (no further processing; 2.4 hrs)
  - Plus remaining tasks (1.7 hrs)

Processing Steps
- Index content for thematic access:
  - Assign one or several descriptors (controlled vocabulary)
- Anonymise and blacken:
  - Text data: personal data (names, dates, or geolocations), translation texts, mouthings and glosses
  - Video data: hands, mouth, or both (black rectangles)
- Translation into English
- Editorial steps:
  - Check spelling and consistency of glosses and mouthings.
  - Crosscheck translations and glosses/mouthings
  - Fix segmentation inconsistencies (overlapping tags)
  - ➡ Iterated reviewing: continuous improvement of annotation quality
  - ➡ Persistent identifiers provided for each transcript file

Conditions of Use
- Protection of privacy of informants and persons mentioned in dialogues
- Small community
  - ➡ Easy identification
- Restriction of metadata to rough categories: age group, sex, geographic region
- Licenses more restrictive than for other projects
- More permissive licenses available upon request

Three Different Formats

"Meine DGS"

- Targets German audience
- Low-threshold access to content of conversations
- Filters for regions, age groups, discussion formats, and topics
- Open archive
  - language, culture, and history of Deaf German individuals
  - Deaf heritage: experiences and attitudes of Deaf individuals
- Usable for e.g. practice purposes (learners and teachers)
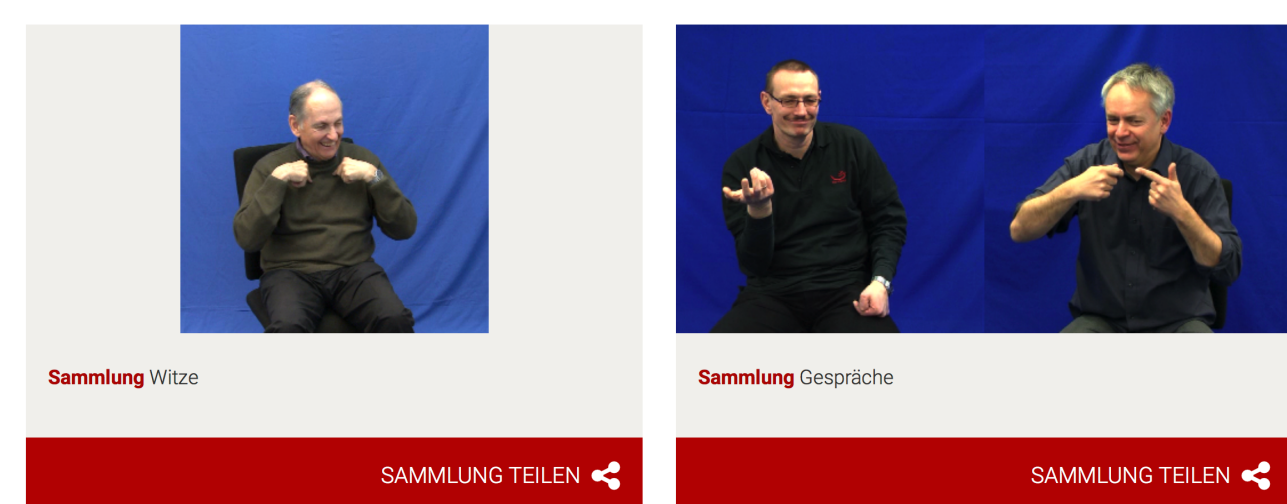
Appearance
- Home page [see figure above]:
  - options for jokes, tasks, or region (map, not on the figure)
- Individual pages (videos) [see figure below]:
  - Videos of signed conversations with German subtitles
  - Subtitles: can be turned off and on at will
  - Additional information (region, dialogue format, topics)
  - Mistake? button: non-public indication of mistakes
  - Share function: for social networks and platforms
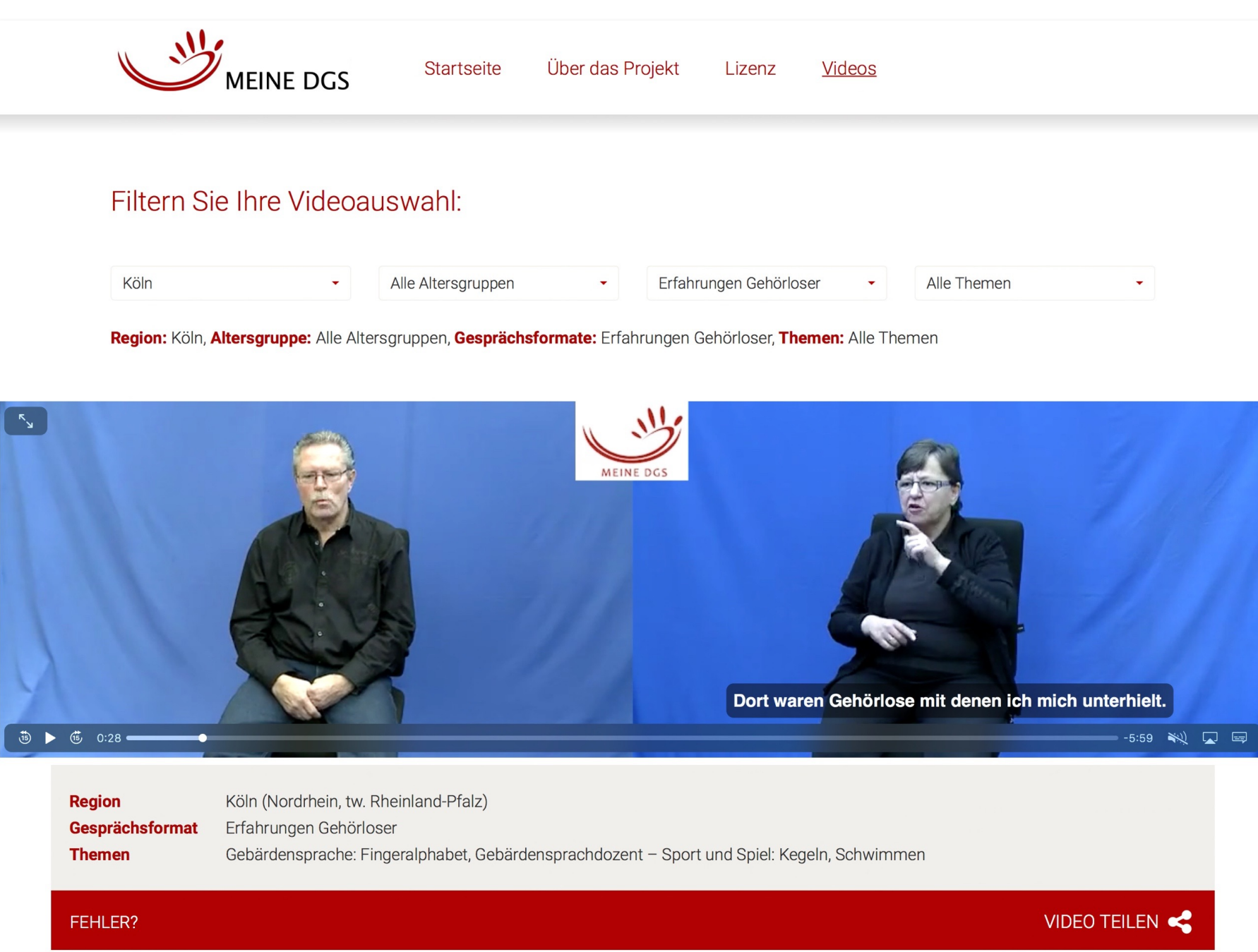
Involving the Language Community
✓ Introduction in DGS
✓ Intuitively usable: not text-intensive, clearly structured, only two page levels, mainly visually oriented, clickable pictures
✓ Mistake? button: attractiveness through interactivity, improvement of the quality of the published data
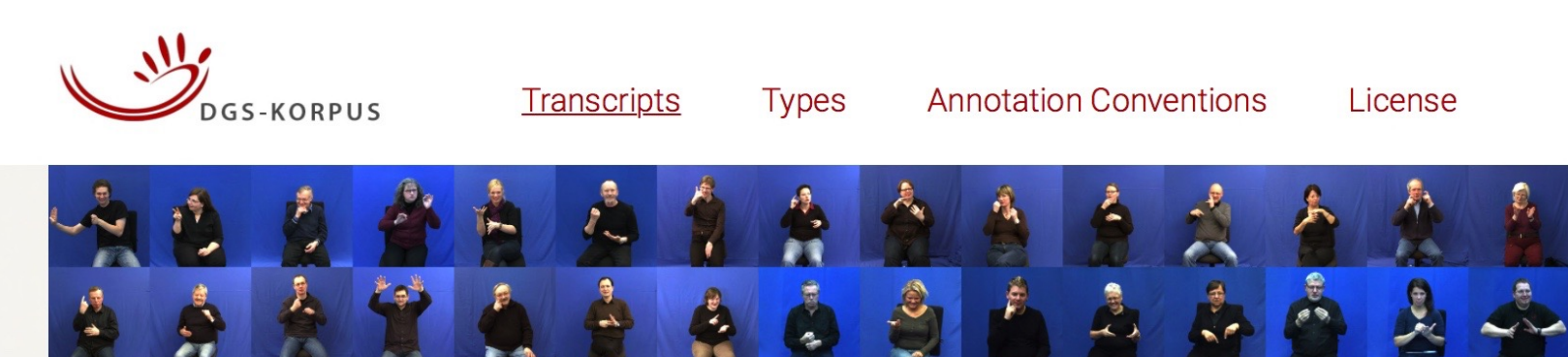✓ Share function: interactive exchange, discussion about specific uses of the data

Comment Function
- Idea: implementation of a comment function
- Supported by members of the focus group
- But: risk of criticizing or insulting informants by anonymous users. Moderator-controlled platform? Costly in resources!
  - ➡ Sufficient solution for the time being: Community building by Mistake? button and Share function

With this format we place importance on a low-threshold entry point to the content of DGS conversations and narrations, especially for the Deaf Community.

Research Portal

- Targets international audience (English only)
- Videos without subtitles, but augmented by annotations:
  - German/English translations, German/English glosses, HamNoSys notations and mouthings (German)
  - Switching between German or English annotations
- Annotation conventions provided
- Types list (over all transcripts)
- Download videos and transcript files (ELAN and iLex format)


Public Corpus Data

Home Page
- List of transcripts, types, annotation conventions, and conditions of use
- List of transcripts: tabular overview over transcripts with information about region (coded in transcript name), age group, sex, format, topics and downloads
- Downloads:
  - More extensive annotation files
  - Additional German/English translations and glosses, HamNoSys notations (citation form of types)
  - Video files (h.264 codec, 640x360, 50 fps) for both informants (frontal views) as well as side views and moderator

Online Transcript View
- Vertical annotation tiers grid with top-down timeline
- Three annotation tiers per informant: *Translation*, *Lexeme/ Sign*, and *Mouthing/Mouth Gesture*
- One tier for moderator: translation (or sum-up)
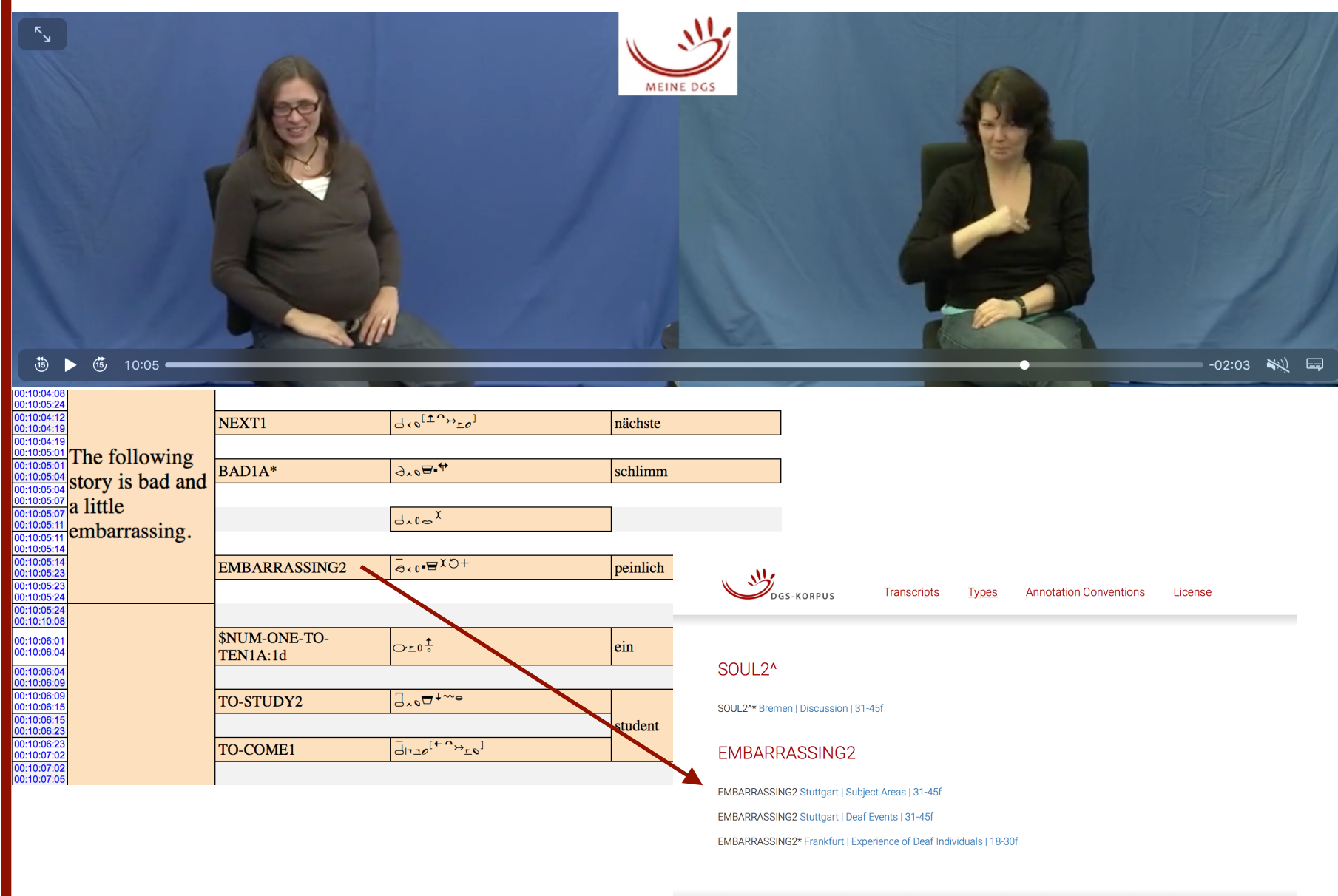
Clickable Glosses
- List of all tokens of the respective type and/or subtype over all transcripts with additional information about region, discussion format, age group, sex

With this format we enable interested users to concentrate on both the content of signed conversations and narrations as well as the language DGS itself.

ANNIS

- Targets international audience (with a background in corpus/ sign language linguistics)
- Online or local version (download)
- Videos in three perspectives (both informants and total view), same annotations as in the Research Portal
- ANNIS: platform-independent open-source search and visualization tool, storage and search possibility for complex corpora with multiple layers
  - ANNIS Query Language (AQL) incl. tutorial and examples
  - Combined searches for annotations and metadata
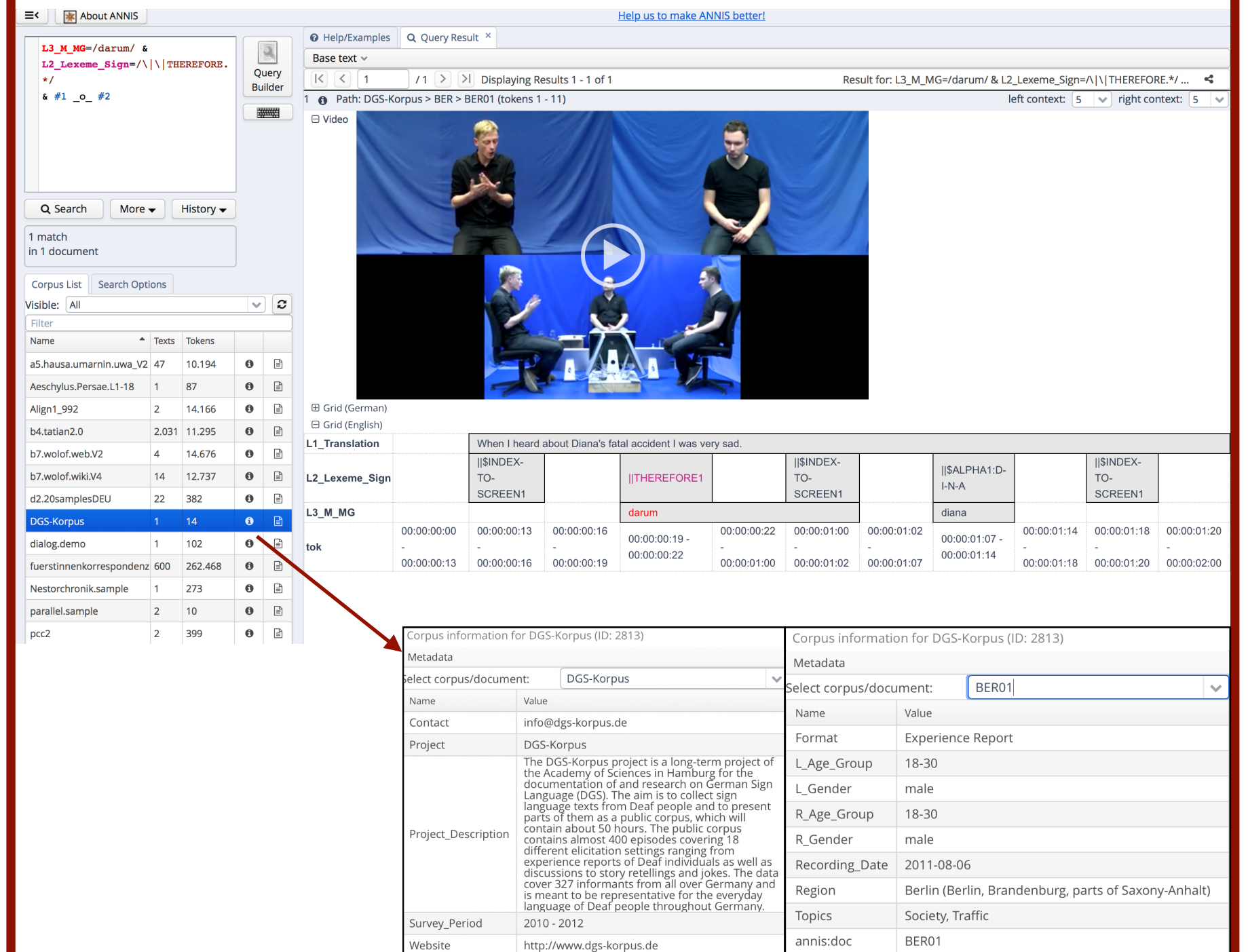

Home page
- Two stationary elements
  - Search field for AQL queries
  - List of publicly accessible corpora
    - Icon leading to metadata information

Presentation of Search Results
- KWIC grid
- Variable size of co-text ("context" preset to five tokens both left and right)
- Timeline similar to Research Portal
- Aligned annotations

ANNIS Query Language
- Search: entries, metadata, sequences, hierarchical orders
- Operators: in-/direct precedence, in-/direct neighbourhood, in-/direct dominance and (identical) overlaps
- Negation of all values and metadata, regular expressions
- Query Builder
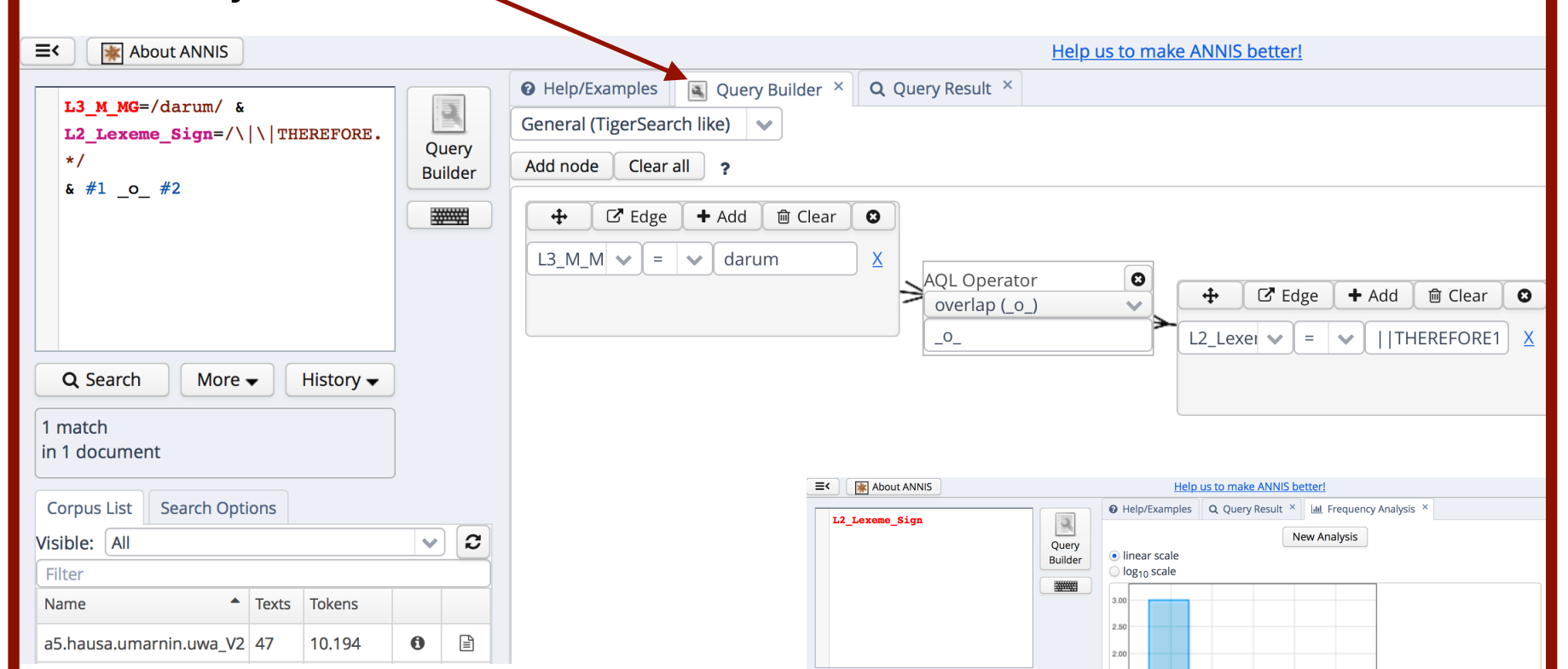
Processing Search Results
- Downloadable in various formats
- Frequency analysis
- Sharing results via link (URL)

With this format we aim to address researchers engaged in corpus based research that might be familiar with ANNIS, similar tools, or query languages, but not necessarily with tools used in sign language corpus linguistics.