Publishing DGS corpus data: Different Formats for Different Needs

Elena Jahn, Reiner Konrad, Gabriele Langer, Sven Wagner, Thomas Hanke
Institute of German Sign Language and Communication of the Deaf, University of Hamburg
Gorch-Fock-Wall 7, 20354 Hamburg, Germany
E-mail: {elena.jahn, reiner.konrad, gabriele.langer, sven.wagner, thomas.hanke}@uni-hamburg.de

Abstract
In 2010-2012, the DGS-Korpus project collected a large corpus of German Sign Language (DGS). Now, a substantial subset of the data is published, namely the Public DGS Corpus. We describe the considerations and decisions taken regarding what part of the data is to be made public, the necessary quality assurance measures to the data preparation as well as the formats of the published data. The corpus is published in three different ways in order to fulfil the needs of a variety of different users. First of all, the data is made available to the language community whose members allowed us to share their recorded language. In addition, we hope that a large number of non-scientific users with various backgrounds will find the data useful. Last but not least, we aim to make the data attractive for users with a scientific background and provide the possibility to conduct studies based on it, irrespective of whether they are familiar with DGS or not.

Keywords: DGS (German Sign Language), corpus building, involvement of the language community, long-term accessibility of sign language data

1. Introduction

In the last ten years, the number of large-scale sign language corpus projects has been growing, in line with the understanding that corpora should form the underpinnings for many research areas, one of them being lexicography. At the same time, there is an increased awareness that the respective language community should benefit from the collected data. On the other hand, funding of corpus work has often been related to specific research questions and not the general usefulness of the data. Large, representtative sign language corpora only recently started to emerge. Thus, it remains a key issue for any sign language corpus work to make the data accessible. Ideally, published corpus data are used frequently by different kinds of users. Therefore, a low-threshold access to the data, suitable for non-scientific users as well as users with a scientific background, should be a requirement.

For sign language corpora, publishing and making data publicly accessible is extremely challenging for ethical reasons (complete anonymization of the informant is not possible in video data), technical reasons (storing video data and keeping them technically up-to-date), historical reasons (the lack of standardised procedures well accepted in the community) and due to a matter of resources (personal and financial) and sustainability.

As with any minority language research, the success of a sign language corpus project strongly depends on the participation and involvement of the language community. Not only are members of the Deaf Community needed in order to gain samples of natural signing by a native signer, but also is the expertise of native signers needed in the process of reviewing translations and annotations. Corpus-based research on sign languages is thus impossible without the help of the Deaf Community. In acknowledgement of the Deaf Community's contribution they should be given continuing access to the data even beyond the period of data collection and processing.

In addition, it is also important that linguistically motivated research on sign languages is facilitated by means of providing corpus data that is suited for publication. However, the detailed exploration of a sign language on basis of a sufficiently large corpus hinges on technical requirements. Therefore, corpus based research on sign languages is a relatively young area of research where the scientific community is still striving for standards. For the aforementioned reasons, it cannot be taken for granted that sign language corpora are published at all. However, technical advancements nowadays facilitate the storage and publication of data online and thus enable projects to share not only their results but also their data. This open-access policy has some major advantages: "When data is accessible to other researchers, research outcomes can be checked by colleagues working in the same field; cross-linguistic studies are facilitated because similar data sets can be recorded for additional languages; the creation of new research groups and the work performed by a single researcher (as for dissertation projects) will become easier because part of the data collection effort can be skipped; finally, seeing in which way other data sets have been collected can lead to the gradual improvement in methodologies for the whole field." (Crasborn et al. 2007: 542)

The DGS-Korpus project is a long-term project of the German Academy of Sciences with two goals: building a reference corpus of DGS and compiling a corpus-based dictionary DGS – German. The raw video data, metadata, and annotations are stored in the iLex database (hereafter iLex), an annotation tool and lexical database that was designed as a multi-user application for annotation and lemmatisation of sign language data (Hanke 2002, Hanke/Storz 2008). Basic annotation includes a translation into German, lemmatisation, and annotation of mouthings/mouth gestures. Detailed annotation is concerned with differentiating between morpho-syntactic inflection, modification, and phonological variation as a basis for the lexicographic analysis and description of signs. Data can be retrieved with customised lists, filters, and queries using SQL. Furthermore, map functions and graphs are integrated, so that e.g. regional distribution of sign variants or variation between age groups can be visually displayed (see Hanke et al. 2017, Langer et al. 2018). Upon request, acces to the corpus data in iLex (as well as the software) is available to researchers outside the project.

Out of the 560 hours of DGS collected in 2010-2012 (Nishio et al. 2010), a subset of about 50 hours is made available as the Public DGS Corpus. It contains almost 400 episodes covering 18 different elicitation tasks ranging from experience reports of Deaf individuals to discussions, story retellings and jokes (see section 2).

We assume that different user groups will address the published data with different expectations. Deaf individuals may be interested in seeing their grandparent generation talk about earlier times, hearing learners of DGS may want to see signs in context or different styles of signing, sign language instructors may search for course material, interpreters for regional variants or DGS equivalents to technical terms, and linguists for appropriate natural DGS signing to conduct crosslinguistic research. Users should be given the possibility to utilise the data as a valuable basis for the investigation of many different questions concerning both the language itself and the language community. To address the different needs and interests, the data is made available via three formats: meine-dgs.de (see 3.1), the Research Portal (see 3.2), and ANNIS (see 3.3).

Providing different formats to access the Public DGS Corpus hopefully contributes to inviting many people to utilise the data. This is supported by encouraging the interaction between users and providing the possibility to report annotation mistakes to the DGS-Korpus team (see 4.).

2. Public Corpus Content

The Public DGS Corpus contains about 50 hours of signed conversations of pairs of interlocutors. The videos are presented bipartitely, with the interlocutors side by side. (In the studio setup, interlocutors were placed facing each other. For more information see Hanke et al. 2010.)

2.1 Prioritising and Selection of Video Material

The videos were carefully selected in order to

- be balanced for region, sex, and age,
- include all elicitation tasks (with the exception of the task "Sign names" for anonymisation reasons),
- cover a great variety of topics,
- cover different styles of signing,
- include each informant at least once.

The corpus shows 327 out of 330 informants (only three informants did not approve the online publication of their data).

The selection process started with a rating of elicitation tasks, in which each project team member rated each task with respect to its importance for the deaf community. As a result, eight tasks were prioritised (in descending order): "Experiences as a Deaf person", "Joke", "Free conversation", "Discussion", "Subject areas", "Experience reports", "Region of origin" and "Deaf events".

With the exception of the task "Joke" (that is, compared to other tasks, rather short), these tasks were proportionally allocated to the planned 50 hours of the public corpus. The remaining tasks were included only exemplarily. We excluded the task "Isolated items" which has a strong lexicographic interest (variation) from the public corpus. In sum, over 47 hours of the videos are selected from the seven tasks listed above, 1.7 hours from remaining tasks, and 2.4 hours from "Jokes".

Within some tasks we presented several stimuli to the informants, e.g. in the task "Subject areas" informants were given four different subjects from which they had to chose two for discussion. These parts (hereafter subtasks) were treated as independent units for annotation workflow. In the next step, subtasks from different informants were selected. For each subtask we revised, among other things, whether the content was appropriate for publication, the style of signing was comprehensible, the video was pleasant to look at, or whether technical difficulties occured during post-production.

2.2 Processing Steps

2.2.1 Indexing Content for Thematic Access

In order to facilitate a thematic access to the videos each selected subtask from the prioritised task list (see above) was indexed for content. A subtask could have one or several descriptors assigned to, but the majority of subtasks were indexed for several descriptors. The descriptors constituted a controlled vocabulary list of about 530 items. Each of these descriptors was assigned to one (or several) of 35 topics. These topics are an extended version of the originally 26 subject areas that were targeted in an elicitation task specifically designed to cover the basic vocabulary of DGS. On the website meine-dgs.de the videos can be filtered by chosing one of the topics (button "Alle Themen"), the more specific descriptors are then displayed below the video screen to facilite a more precise selection according to interest of the user. In the Research Portal the topics are listed under the column "Topics".

2.2.2 Translation into English Version

With the exception of the task "Joke" all selected subtasks were translated into German and lemmatised as part of the basic annotation. In a second step, they were translated from German into English. This enables researchers knowing neither DGS nor German to browse the content of the public corpus (in the research portal and in ANNIS). In addition, the German glosses were also translated into English and are displayed in the English version of the online transcript view.

2.2.3 Blackening and Anonymisation

We spent some effort to anonymise parts of the signing that should be exempt from the online publication. If stretches to be anonymised were too long, the subtask was not selected for inclusion in the Public DGS Corpus. In other cases, we decided to shorten the subtask, mostly at the beginning or the end. Finally, we have some cases left where stretches had to be blackened within a subtask (in general only a few seconds). In order to anonymise personal data of the informants or third persons (names, dates like birthday, or geolocations) we tagged these sequences, decided whether hands, mouth, or both had to be blackened and generated rectangle coordinates as annotations. These coordinates had to be checked manually in the frontal and profile view of the informants. When exporting the movie files the designated blocks were rendered black. Besides the videos, also translation texts, mouthing annotations and glosses hat do be identified and processed in order to produce anonymised texts and annotations (for details see Bleicken et al. 2016).

2.2.4 Editorial Steps

The publication of a sign language corpus requires additional steps not crucially necessary to work with the data in-house.

A built-in spell checker in iLex (for German and English) supported the annotators when aligning the German translations. Glosses and mouthings were checked manually.

Further on, we checked translations against lemmatisation and mouthings in order to reach a high consistency of the

annotations. This checking helped to fill translation gaps and revise unclear passages or to correct token-type mismatches and mouthings. The experience of the Deaf team members, Deaf students, and CODAs was indispensable and most valuable in this step.
Also, inconsistencies in the segmentation of subtasks, translations, tokens and mouthings/mouth gestures had to be checked, e.g. a translation tag should not start before or end after a subtask tag, a translation tag should not start or end in between a token or mouthing/mouth gesture tag. Overlapping translation tags of informant A and B with no significant signing had to be corrected.
Last but not least, each processing step helps to improve the quality of the annotations. Annotators comment and give feedback to translation and lemmatisation that were reviewed. This checking procedure has the drawback that tags were changed or comments were added after the corresponding annotation or checking step was already done. But this seems to be unavoidable when working with a team of 15 colleagues and over 30 student co-workers.

2.2.5 Persistent Identifier

We provide persistent identifiers for individual transcripts to make them quotable in a revision-savvy way.

3. Different Formats for Different Needs

The formats in which the Public DGS Corpus is distributed are the following:

- The website meine-dgs.de is a low-threshold access to the data. In this portal, videos are presented together with German translations as subtitles. Here, the focus is on content-related access.
- The Research Portal provides the video data with basic annotations as well as metadata on the informants for linguistic and related research. Annotation data (in German and English) is made available for download in ELAN and iLex format, or can be previewed in the web browser.
- In order to also provide easy access for researchers not familiar with annotation environments prevalent in sign language research, but with corpus tools in general, we also plan to make our data accessible via ANNIS (ANNotation of Information Structure; http://corpus-tools.org/annis/; Krause & Zeldes, 2016). ANNIS is a corpus query tool for visualization and querying multi-layer corpus data that comes along with its own query language.

3.1 meine-dgs.de

The first publication format, meine-dgs.de, is a website where users can watch the signed conversations or narratives with subtitles showing the translations into German, except jokes. In addition to the main page with the videos, the website contains information about the project, license terms and a page where the videos can be filtered for region, age groups, dialogue formats and main topics.
The website meine-dgs.de is meant to address users that are interested in the content of the conversations and narratives. It provides a low-threshold access to the data and is thus suitable for both users without a scientific background and users with a scientific background that would like to get familiar with the data. Also, users with a scientific background that is not linguistics or sign languages might find the data interesting, e.g. for studies concerning Deaf Culture or the way in which Deaf individuals have experienced decisive events. DGS is known to have regional variants, therefore users might want to search for videos from specific regions only.
The appearance of the website is as follows. On the main page, users can decide for jokes only ("Sammlung Witze" leading on a page with the format "Witze" (88 jokes) preselected), for all subtasks ("Sammlung Gespräche" with no format preselected), and the possibility to preselect the region via a map ("Sammlung Regionen"). For each video a short description is provided which contains information about the region (city or geographical area) where the conversation has been filmed (and the interlocutors are rooted), the dialogue format and the topics. This information is meant to help the user to get an overview over the data and select the most interesting videos.
The video contains subtitles that can be turned off and on at will. Below the video on the left, a mistake button ("FEHLER?") is implemented that allows for a non-public indication of mistakes to the DGS-Korpus team. On the right side a share button ("VIDEO TEILEN") enables to share the respective video in various social networks and platforms.

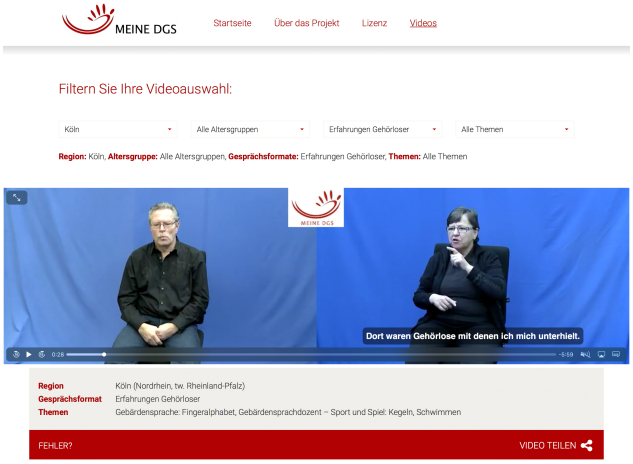


Figure 1 : meine-dgs.de

The general aim of the publication of the data on the website meine-dgs.de was to allow the user to concentrate on the content of the signed conversations or narratives. With this format we place importance on a low-threshold entry point for any interested person. Also, we hope to provide valuable data for learners and teachers of DGS who might use the videos for practice purposes. Researchers that are interested in getting an overview of the content of the conversations and the recording situation might find the site helpful, too. Also, the website serves as an open archive for language, culture, and history of Deaf individuals.

3.2 Research Portal

This portal is made for users with a scientific background who are interested in the content of the conversations and narratives, but with a focus on the language DGS itself.

Like meine-dgs.de, the Research Portal is accessible without prior registration. As it is supposed to address an international audience, the website is in English. It provides the same videos, here without subtitles, but augmented by annotations. It starts with a list of "Transcripts" (i.e. the subtasks), offers a "Types" list with all types used for lemmatising the tokens in the public corpus, links to the "Annotation Conventions" and informs about the conditions of use ("License"; see 5.). In the header a banner displays all informants.

The body shows a list of all subtasks. Instead of filters the subtasks are listed by the transcript name like "dgskorpus_ber_01" coding the region (ber=Berlin) and a running number for the elicitation session. Further codes are: fra (Frankfurt), goe (Göttingen), hb (Bremen), hh (Hamburg), koe (Köln), lei (Leipzig), mst (Münster), mue (München), mvp (Mecklenburg-Vorpommern), nue (Nürnberg), sh (Schleswig-Holstein), and stu (Stuttgart). Thus, the filter "Region" is dispensable. Age group, format, and topics are further columns in this list. The next columns contain icons to download annotation and video files. Annotation files are offered for iLex and ELAN import and are more extensive than the online transcript as they include both German and English translations and glosses, and additionally HamNoSys notations of the citation form of the types. Video files (h.264 codec, 640x360, 50 fps) are provided not only for informant A and B, but also for a total perspective with both informants in profile view and the moderator in the middle.

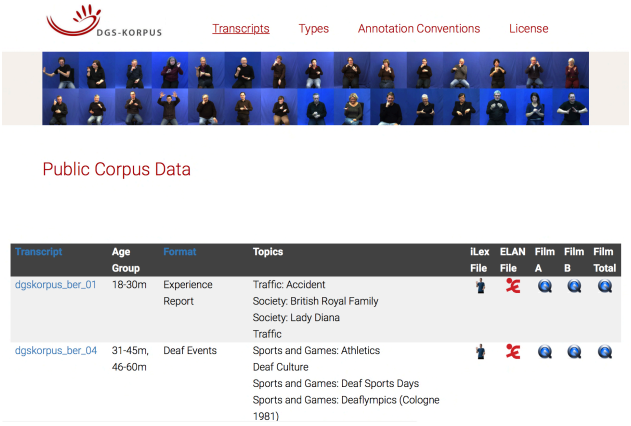


Figure 2 : Research Portal main page

By clicking on the "Transcript" name, videos and annotations can be browsed by an online transcript view (with the possibility to switch between a German and an English version). In this way, it differs from other access formats to sign language corpora. The online transcripts may be of interest also for users without a scientific background. It gives everyone a glimpse on how basic research in sign language corpus linguistics looks like and makes the results of our work transparent.

In the online transcript view, the videos with both informants are displayed at the top, with the transcript beneath. The annotation tiers are arranged in a vertical grid with a top-down timeline (as opposed to a horizontal grid many researchers may be used to). The timeline shows timecode start and end for each tag.[1] Three annotation tiers for each informant exist: Translation, Lexeme/Sign, and Mouthing/Mouth Gesture. Just like the video screen, the tiers of informant B are on the left, those for informant A on the right side. It is not very often that the moderator interacts. Therefore, we skipped the total perspective in the online view and added a seventh tier for a summary of the moderator's interaction. To keep the tiers apart, they have different background colours for informants and moderator. A link allows switching to the German version (with translation and glosses in German).

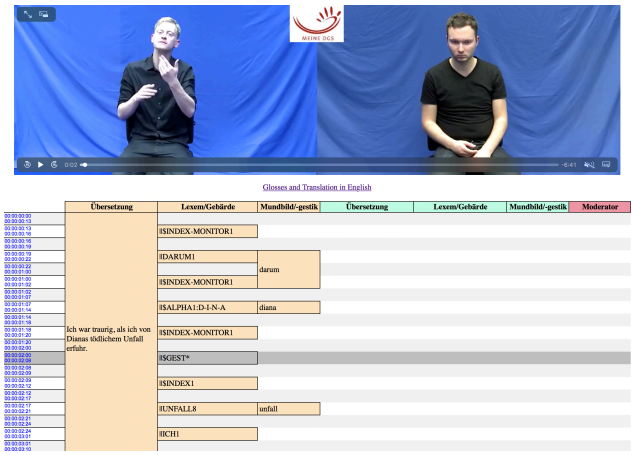


Figure 3 : Research Portal Annotation Tiers

3.2.1 Annotation Tiers

The German translation should be as close to the DGS utterance as possible. We did not aim for a free translation, because the translation should guide the mostly hearing student annotators. Contracted sign language interpreters conducted a first translation. The student co-workers splitted and time-aligned these texts into 'sentence'-like utterances. As Johnston (2016: 14) posits, these "translation sentences are not attempts to segment the [DGS] text into its potential language-specific syntactic or grammatical units". They are searchable and define preliminary utterance units when looking for the context of a sign token. The translation into English is a free translation. Its purpose is to give access to the content of the DGS videos to those knowing neither DGS nor German.

Mouthings are very frequent in DGS. They are an important clue to the meaning of a DGS sign token which, in combination with the sign form, can be used to search for the appropriate type the token should be matched to. Thus, we decided to also annotate mouthings in the phase of basic annotation. Mouthings are annotated in lower case to make them distinct from German words. As we focussed on the meaning of the mouthed word and not its actual articulation, at least the intended word (word stem) to be lip-read should be annotated. Incomplete mouthings are supplemented (in curly brackets), uncertainties are

[1] For performance reasons the videos have a framerate of 25 fps, the timeline instead follows a 50 fps rate to be consistent with the timecodes in the ELAN and iLex import files. As a consequence, the videos in the online view are not suitable for frame-to-frame inspection. For this, one has to use the download files.

marked by "??". As mouthings in DGS refer to German words, the articulation features are different from e.g. mouthed English words. We therefore do not provide a translation of mouthings.
Mouth gestures are movements of the mouth region with no connection to words of the vocal language. With a focus on lexical signs, we did not aim for classifying mouth gestures by form features. They are annotated in a simplistic way by just adding "[MG]" in the Mouthing/Mouth Gesture tier.
The annotation files are complemented by two HamNoSys tiers with notations of the citation form of types in the Lexeme/Sign tiers that are available after download. Annotation Conventions for the Lexeme/Sign tiers are explained in the following.

3.2.2 Annotation Conventions

There are two main aspects in which our approach differs from those of other sign language corpus projects: the role of mouthings which led us to implement a type hierarchy (double glossing) in the database model, and double-token tags in the token tier instead of separate gloss tiers for left and right hand.

3.2.2.1 Type hierarchy (double glossing)

In brief, we are convinced that following the principle of idiomaticity does not fit the needs of an adequate description of a sign language lexicon. The reason why (lexical) signs can cover a far wider range of meanings than words is iconicity. Sign languages exploit the possibilities to express the visually perceivable world in a visual-gestural modality which also allows for integrating words of the surrounding vocal language by way of mouthings. Conventionalisation should not only be applied for distinguishing lexical from productive signs, but also for sign-mouthing combinations (for further details see König et al. 2008, 2010, Konrad et al. 2012).
Glosses in the "Lexeme/Sign" tier refer either to a type or a subtype. Types correspond to lexical entries which have at least on conventionalised meaning. In order to group these form-meaning combinations, often expressed by conventionalised sign-mouthing combinations, we use subtypes. Each type (parent) has at least on subtype (child). Tokens of conventional sign-mouthing combinations are matched to the appropriate subtype, tokens of productive sign-mouthing combinations are matched to the type. This kind of pre-sorting supports the lexical description of sign types.
Glosses are labels for sign types/subtypes representing unique type entities in the lexical database and can be taken as ID-glosses, regardless whether in German or English (Johnston 2008; Konrad/Langer 2009). Glosses at the type level are marked by a superscript after the gloss name as e.g. FLACH1^ (PLANE1^). One of its subtypes is TISCH1 (TABLE1), without superscript. In iLex we annotate form deviation to the token tag and sort tokens for morpho-syntactic patterns or modification by using qualified types (see Konrad et al. 2012). In the Research Portal we only show types and subtypes. Tokens that differ from the types citation form are marked by an asterisk after the gloss name, e.g. TABLE1*.

The online view of transcripts not only allows to browse the annotations, but also can be used to list all tokens of a type and subtype lemmatised in the whole public corpus. By clicking on the gloss name in the "Lexeme/Sign" tier a new page opens with all the tokens that are matched to the corresponding type and/or subtype, irrespective whether the type or subtype gloss is clicked. In addition to the gloss name, several metadata are provided: region, format, age group, and sex. The following screenshot shows the tokens matched to the type SOUL2^ and the subtype EMBARRASING2:

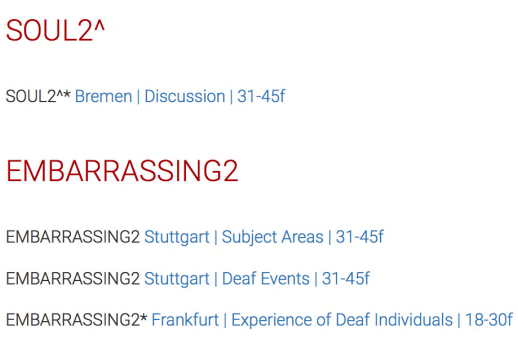


Figure 4 : Listing of Tokens from Types and Subtypes

3.2.2.2 Double Tokens

Many researchers using e.g. ELAN as annotation tool have two token tiers, one for each hand. Two-handed signs are lemmatised by annotating the same gloss in each tier. In order to make the annotation easier and less time-consuming we opted for one token tier which allows for annotating one type for each hand. Two-handed signs are either annotated in the right or left hand slot: For asymmetric signs the slot of the active hand is used. For symmetric signs the right hand slot is used as a default.
A sign articulated with the right hand – being either a one- or two-handed sign – is displayed in the type-/subtype-gloss tier by one gloss. If the sign is articulated with the left hand, the gloss is preceded by a double bar, e.g. ||HAUS1A (HOUSE1A). A complex sign construction shows two glosses separated by a double bar, e.g. OMA2||$INDEX1 (GRANDMA1||$INDEX).

3.2.2.3 Glossing conventions

Although basic annotation of sign language texts should be as theory-neutral as possible, it cannot do without any theoretical assumptions. One is the distinction of three sign categories: lexical signs (cf. Johnston 2016: fully-lexical signs), productive signs (cf. Johnston 2016: partly-lexical signs), and others (cf. Johnston 2016: non-lexical signs). In the following we just mention some of the glossing conventions, for a detailed description see "Annotation Conventions" in the Research Portal.
Lexical signs are glossed by German (English) words. Different numbers are used to group lexical variants, e.g. FRAU4 (WOMAN4) and FRAU5 (WOMAN5). Phonological variants are grouped together by using the same gloss name and number followed by different letters, e.g. FRAU2A (WOMAN2A) and FRAU2B (WOMAN2B). Productive signs are glossed as $MAN (abbreviation for "manual activity"; $PROD for "productive sign"). For

grouping together type categories in a sorted type list, we use prefixes like $NAME- (name signes), $INDEX (pointing signs), $ALPHA (fingerspelling), or $GEST (gestures).

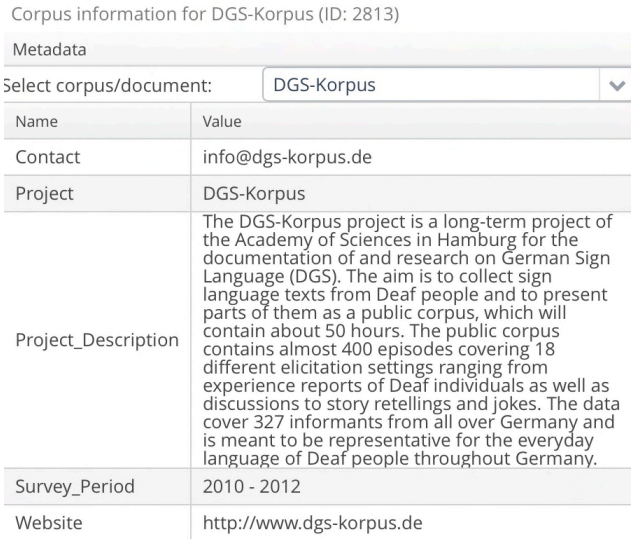
3.3 ANNIS

The platform-independent open-source search and visualization tool ANNIS comes along as both a web-application and a local version. ANNIS was put forth by a DFG project, the SFB632 “Information Structure: The Linguistic Means for Structuring Utterances, Sentences and Texts", realised by researchers of the University of Potsdam, the Humboldt-University of Berlin and the Free University of Berlin. While the project ended in 2015, ANNIS has been used by further projects ever since. It is meant to be a storage and search possibility for complex corpora with multiple layers that can originate from different annotation tools. Along with the growing number of multimodal corpora, ANNIS allows to implement video data as well as linking parts of a video with the associated annotations. It also enables users to directly search for annotations with the ANNIS query language (AQL; for more information see Rosenfeld 2010), that provides powerful search options. With every corpus that is published in ANNIS, search examples in AQL are provided, either automatically or preset by the researcher. Clicking on these example queries leads to their results. AQL allows searching for, inter alia, entries and metadata, sequences and hierarchical orders. Complex searches can be formulated in AQL, too, in accordance with the following scheme. First, one or more attribute-value pairs are defined. Second, the relationships between the nodes are defined, using among others the following operators: in-/direct precedence, in-/direct neighbourhood, in-/direct dominance and (identical) overlaps. Regular expressions can be used, too. All values can be negated and so can metadata. Search results can be displayed in different views, like syntax trees or dependency relation schemes. Since annotation tiers are not hierarchically linked in the Public DGS Corpus, results are presented in a KWIC (key word in context) table view, called grid. The size of the context is preset to five tokens both left and right of the search result (but can be varied). Once a search is successfully carried out, a frequency analysis on the search results can be conducted. For further statistical or other analyses, results can be downloaded in various formats. Results of a search or a frequency analysis can be shared via a link.

ANNIS was choosen as a third presentation format in order to enable users to directly search the data online without the need to register, download data, install new programs and learn a completely new query language. The essential features of the ANNIS query language might be familiar to most researchers engaged in corpus based research. Since many researchers might already be used to ANNIS or similar corpus search tools, the Public DGS Corpus in ANNIS is therefore mainly meant to address those researchers. Nevertheless, also users without a (corpus) linguistic background can find an easy access-point to a scientifically motivated approach to the data with ANNIS.

Using ANNIS requires getting familiar with the tool and its query language. Also, ANNIS is not meant to be another content-related access point. Users should at least roughly know the content, the metadata and the annotation conventions used. Watching the complete video collection in ANNIS will most likely be uncomfortable – for this matter, meine-dgs.de is more advisable. While ANNIS also provides the possibility to store corpora in a restricted area, to which only researchers are granted access after registration with a university e-mail address, the Public DGS Corpus will be released in the public area, in which no prior registration is needed.

The Public DGS Corpus is presented in ANNIS as follows. Both in the online and the local version, the ANNIS main page contains two stationary elements, namely a box where AQL queries can be typed in and a list of publicly accessible corpora. (On a fixed tab, a help page and a tutorial can be opened at any time). Each corpus is listed with a small icon leading to metadata information about the corpus.


Corpus information for DGS-Korpus (ID: 2813)

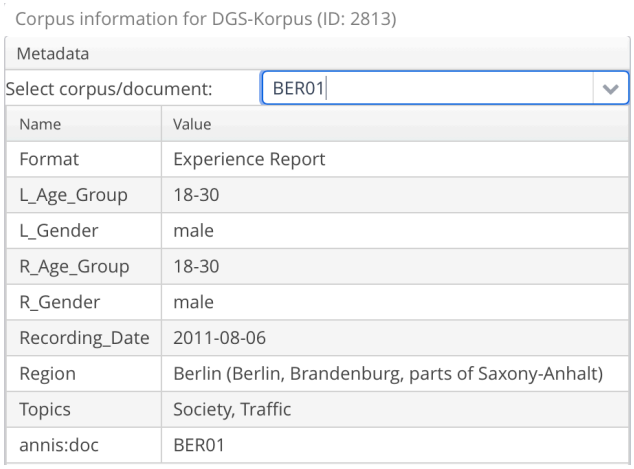
Metadata

Select corpus/document: DGS-Korpus

Name	Value
Contact	info@dgs-korpus.de
Project	DGS-Korpus
Project_Description	The DGS-Korpus project is a long-term project of the Academy of Sciences in Hamburg for the documentation of and research on German Sign Language (DGS). The aim is to collect sign language texts from Deaf people and to present parts of them as a public corpus, which will contain about 50 hours. The public corpus contains almost 400 episodes covering 18 different elicitation settings ranging from experience reports of Deaf individuals as well as discussions to story retellings and jokes. The data cover 327 informants from all over Germany and is meant to be representative for the everyday language of Deaf people throughout Germany.
Survey_Period	2010 - 2012
Website	http://www.dgs-korpus.de

Figure 5 : Public DGS Corpus Metadata in ANNIS

Metadata can also be added for individual documents. Thus, the document metadata can be used as values in queries.


Corpus information for DGS-Korpus (ID: 2813)

Metadata

Select corpus/document: BER01

Name	Value
Format	Experience Report
L_Age_Group	18-30
L_Gender	male
R_Age_Group	18-30
R_Gender	male
Recording_Date	2011-08-06
Region	Berlin (Berlin, Brandenburg, parts of Saxony-Anhalt)
Topics	Society, Traffic
annis:doc	BER01

Figure 6 : Document Metadata in ANNIS

Selecting the Public DGS Corpus leads to a list of example queries. Clicking on an example query leads to the search result. This is a user-friendly access to the data and gives a good first impression of the query language.
Search results are presented by means of two grids (for English and German annotations) that can be folded up and out at will. The video is displayed above. The grids contain the same tiers that are displayed on the Research Portal online view, namely three tiers per informant (Translation, Lexeme/Sign and Mouthing/Mouth Gesture) and one for the moderator.

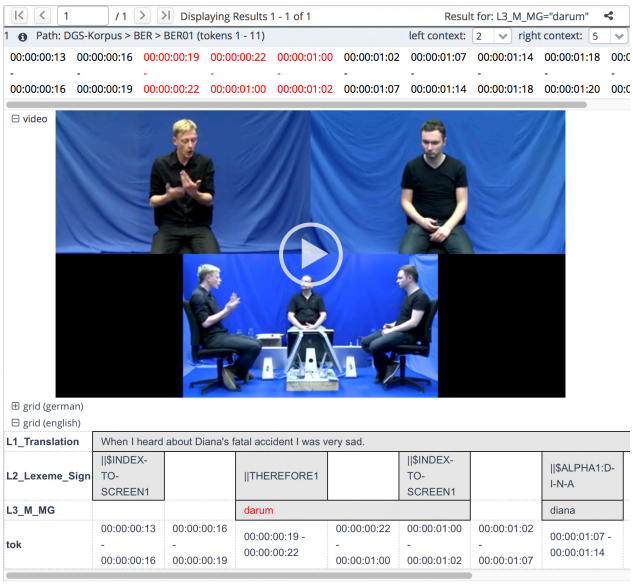


Figure 7 : Presentation of Results in ANNIS

With the presentation of the Public DGS Corpus in ANNIS, we provide access for corpus-based research that is based on a variety of different information such as different layers of annotation and relations between annotation tiers and metadata information. The publication of the public corpus in ANNIS makes the data scientifically usable, facilitates perusing of the data and allows the sharing of search results with other interested parties by means of a handy link.

4. Involving the Language Community

meine-dgs.de is a low-threshold website that provides an easy access point. For users from the Language Community, whose first language is a visual language and who therefore might feel more natural with signed information, we provide information about the low-threshold format meine-dgs.de by means of a signed video introduction. Also, meine-dgs.de is designed to be intuitively usable. It is not text-intensive, clearly structured, and in general mainly visually oriented, with clickable pictures and short access paths.
Furthermore, we included features that facilitate interactivity and the involvement of the Language Community, namely the "Mistake" button and the share function.

As for the "mistake" button, interactivity makes the data and its use even more attractive and helps to improve the quality of the published data. As described above, published data has gone through a process of reviewing and examining. Nevertheless, mistakes can never be completely avoided.
An interactive exchange and the establishment of a discussion about the intrinsic value of the data for specific use cases could increase the users' interest in the data. Members of the focus group, a group of informants that are well rooted in the Deaf Community, supported this idea. Although the increase of interactivity through a comment function would be a great advantage, we are also well aware that it is difficult to filter comments and sort out offensive, nonsensical or other undesired comments. While this is usually a typical and tolerated sideeffect of online platforms with comment functions, we aim to strictly avoid situations, in which anonymous users criticise or insult informants. The benefits and costs of a moderator-controlled platform must be weighed. Up to now, it is only possible to share videos from meine-dgs.de to other platforms and social networks. This allows users to draw attention to videos they find especially interesting or valuable and also enables users to get in contact with each other.
For the moment, we take this as a sufficient solution to initiate the building of a community, which will be observed and inspected from time to time, in order to detect a good moment to organise an interaction directly on meine-dgs.de.

5. Conditions of Use

Obviously, publishing video data and at the same time protecting the rights of the informants is more difficult than publishing data collections that consist of texts or audio files. The privacy of the informants themselves as well as all persons mentioned in the dialogues has to be respected. Since the sign language community is a relatively small community, small hints on the identity of third persons mentioned might be enough for identification. For these reasons, we exclude data from publication when in doubt and restrict the publication of metadata to very rough categories, age group, sex, and larger geographic region.
We also attach great importance to matters of ethics and therefore follow the wishes of the informants how their data can be used. As they need to cover all data in the public corpus, the licenses for using the data are therefore more restrictive than for some other projects. More permissive licenses are available only upon request.

6. Acknowledgements

This publication has 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 German Academies of Sciences and Humanities.

7. Bibliographical References

Bleicken, Julian / Hanke, Thomas / Salden, Uta / Wagner, Sven. (2016). Using a Language Technology Infrastructure for German in order to Anonymize German Sign Language Corpus Data. In: Calzolari, Nicoletta et. al. (eds.): *LREC 2016 Proceedings. Tenth International Conference on Language Resources and Evaluation (LREC), May 23-28, 2016, Portorož, Slovenia.* ELRA. pp. 3303-3306. [URL: http://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/LREC2016-419_Paper.pdf; last access 2018-02-23].

Crasborn, Onno A. / Mesch, Johanna / Waters, Dafydd / Nonhebel, Annika / Van der Kooij, Els / Woll, Benice / Bergman, Brita. (2007). Sharing sign language data online: Experiences from the ECHO project. In: *International journal of corpus linguistics*, *12*(4). pp. 535-562.

Hanke, Thomas. 2002: iLex. A tool for Sign Language Lexicography and Corpus Analysis. In: González Rodriguez, Manuel / Paz Suarez Araujo, Carmen (eds.): *Proceedings of the third International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain; Vol. III.* Paris: ELRA. pp. 923-926. [URL: http://www.lrec-conf.org/proceedings/lrec2002/pdf/330.pdf; last access 2018-02-23].

Hanke, Thomas / Storz, Jakob. (2008). iLex – A database tool for integrating sign language corpus linguistics and sign language lexicography. In: *LREC 2008 Workshop Proceedings. W 25: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora.* Paris: ELRA. pp. 64-67. [URL: http://www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf; last access 2018-02-23].

Hanke, Thomas / König, Lutz / Wagner, Sven / Matthes, Silke. (2010). DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In: *P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz, A. Schembri (2010, Eds.) Corpora and Sign Language Technologies. [Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages. 7th International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta.].* Paris. ELRA. pp. 106-109. [URL: http://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/Hanke_et_al_2010_Studio.pdf; last access 2018-02-23].

Hanke, Thomas / Konrad, Reiner / Langer, Gabriele / Müller, Anke / Wähl, Sabrina. (2017). Detecting Regional and Age Variation in a Growing Corpus of DGS. Poster presented at the Workshop: *Corpus- based approaches to sign language linguistics: Into the second decade. July 24, 2017, Birmingham*. [URL: http://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/DGS-Korpus_Poster_Birmingham2017_Variation.pdf; last access 2018-02-23].

Johnston, Trevor. (2016). Auslan Corpus Annotation Guidelines. November 2016 version. [Online resource; URL:http://www.academia.edu/29690332/Auslan_Corpus_Annotation_Guidelines_November_2016_revision].

Johnston, Trevor. 2008: Corpus linguistics and signed languages: no lemmata, no corpus. In: Crasborn, Onno / Efthimiou, Eleni / Hanke, Thomas / Thoutenhoofd, Ernst D. / Zwitserlood, Inge (eds.): *LREC 2008 Workshop Proceedings. W 25: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora.* Paris: ELRA. pp. 82-87. [URL: http://www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf; last access 2018-02-23]

König, Susanne / Konrad, Reiner / Gabriele Langer. (2008). What's in a sign? Theoretical lessons from practical sign language lexicography. In: Quer, Josep (ed.): *Signs of the time. Selected papers from TISLR 2004.* Hamburg: Signum. pp. 379-404.

König, Susanne / Konrad, Reiner / Langer, Gabriele / Nishio, Rie. (2010). How Much Top-Down and Bottom-Up do We Need to Build a Lemmatized Corpus? Poster presented at the Conference: *Theoretical Issues in Sign Language Research Conference (TISLR 10), Sept 30 - Oct 2, 2010, Purdue University, Indiana, USA.* [URL: http://www.sign-lang.uni-hamburg.de/dgs-korpus/tl_files/inhalt_pdf/PosterTISLRTranskription1_R12.pdf; last access 2018-02-23].

Konrad, Reiner / Langer, Gabriele. (2009). Synergies between transcription and lexical database building: The case of German Sign Language (DGS). In: Mahlberg, Michaela / González-Díaz, Victorina / Smith, Catherine (eds.): *Proceedings of the Corpus Linguistics Conference (CL2009), July 20-23, 2009, University of Liverpool, UK.* [URL: http://ucrel.lancs.ac.uk/publications/cl2009/346_FullPaper.doc; last access 2018-02-23].

Konrad, Reiner / Hanke, Thomas / König, Susanne / Langer, Gabriele / Matthes, Silke / Nishio, Rie / Regen, Anja. (2012). From form to function. A database approach to handle lexicon building and spotting token forms in sign languages. In: *O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, J. Mesch (2012, Eds.) Interactions between Corpus and Lexicon [Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages. 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey.]* Paris. ELRA. pp. 87–94. [URL: http://www.lrec-conf.org/proceedings/lrec2012/workshops/24.Proceedings_SignLanguage.pdf; last access 2018-02-23].

Krause, Thomas / Zeldes, Amir (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1). pp. 118-139. [URL: https://doi.org/10.1093/llc/fqu057].

Langer, Gabriele / Müller, Anke / Wähl, Sabrina. (2018). Queries and views in iLex to support corpus-based lexicographic work on German Sign Language (DGS). [this issue].

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. In: *P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz, A. Schembri (2010, Eds.) Corpora and Sign Language Technologies. [Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages. 7th International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta.]* Paris. ELRA. pp. 178-185. [URL: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W13.pdf; last access 2018-02-23].

Rosenfeld, Viktor. (2010). An implementation of the Annis 2 query language. Humboldt University Berlin.