Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 75–82
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Extending the Public DGS Corpus in Size and Depth

Thomas Hanke, Marc Schulder, Reiner Konrad, Elena Jahn
Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Germany
{thomas.hanke, marc.schulder, reiner.konrad, elena.jahn}@uni-hamburg.de

Abstract

In 2018 the DGS-Korpus project published the first full release of the Public DGS Corpus. This event marked a change of focus for the project. While before most attention had been on increasing the size of the corpus, now an increase in its depth became the priority. New data formats were added, corpus annotation conventions were released and OpenPose pose information was published for all transcripts. The community and research portal websites of the corpus also received upgrades, including persistent identifiers, archival copies of previous releases and improvements to their usability on mobile devices. The research portal was enhanced even further, improving its transcript web viewer, adding a KWIC concordance view, introducing cross-references to other linguistic resources of DGS and making its entire interface available in German in addition to English. This article provides an overview of these changes, chronicling the evolution of the Public DGS Corpus from its first release in 2018, through its second release in 2019 until its third release in 2020.

Keywords: German Sign Language (DGS), Linguistic Resource, Corpus, Resource Extension

1. Introduction

For the past eleven years, the DGS-Korpus project (Prillwitz et al., 2008) has been building the ***DGS Corpus***, an annotated collection of dialogues between native signers of German Sign Language (DGS). Based on this corpus, two publicly accessible resources are created by the project:

1. The ***Public DGS Corpus***, a subset of the full project corpus, accessible via two formats:
   (a) ***MY DGS***[1], a community portal for the Deaf community and others interested in DGS, which offers video recordings of selected dialogues with optional German subtitles, and
   (b) ***MY DGS – annotated***[2], a research portal for the international scientific community, which offers an annotated corpus of DGS for linguistic research.
2. ***Digitales Wörterbuch der Deutschen Gebärdensprache* (*DW-DGS*)**, the first corpus-based digital dictionary of DGS–German.

These resources are released and extended progressively throughout the life time of the project. For example, while the first preliminary version of the community portal was released in 2015, its first full release, as well as the first release of the research portal, happened in 2018 (Jahn et al., 2018). The first preliminary release of the *DW-DGS* is scheduled for 2020.

In this article we present how the *Public DGS Corpus* and its two portals have been extended in subsequent releases after 2018. This involves the addition of new data, corrections to the subtitles and annotations, as well as several new features, such as new data formats, body pose information, unique identifiers, cross-references to other resources, collocation views and more. Some of these changes affect both the research and community portal, while others are only of relevance to researchers and therefore limited to the research portal.

The remainder of the article is structured as follows: Section 2 briefly introduces the DGS-Korpus project, its corpus creation efforts and how it has published corpus data up until the first full release of the *Public DGS Corpus* in 2018. The remaining chapters then address the changes introduced in 2019 (Release 2) and 2020 (Release 3). Section 3 presents the different kinds of content that have been added or extended, while Section 4 describes the different data formats in which data can be accessed on the research portal. Section 5 concludes the article by providing an outlook on the future directions that the corpus will take.

2. The *DGS Corpus*

The DGS-Korpus project[3] is a long-term project of the Academy of Sciences and Humanities in Hamburg. It was started in 2009 and aims to build a reference corpus of German Sign Language (DGS), publish a subset of about 50 hours with annotations in both German and English and to compile a corpus-based dictionary DGS – German. From 2010 to 2012 data were collected from 330 informants at 12 different locations in Germany. The selection of informants was balanced for sex, age, and region. The informants were filmed in pairs and presented 20 different elicitation tasks, which cover a broad variety of discussion formats and topics with a focus on dialogue and natural signing (Nishio et al., 2010). For information on the studio set-up, see Hanke et al. (2010).

The footage of the *DGS Corpus* consists of over 1150 hours of recordings, containing about 560 hours of near-natural DGS signing. The project uses iLex[4], 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 and Storz, 2008).

The basic annotation of these videos comprises translation into German, lemmatisation and annotation of

[1] http://meine-dgs.de
[2] http://ling.meine-dgs.de
[3] http://dgs-korpus.de/
[4] www.sign-lang.uni-hamburg.de/ilex/

mouthing/mouth gestures. The translations were carried out by professional interpreters, alignment of these texts and further annotation mainly by student assistants. Lemma revision (Konrad and Langer, 2009; König et al., 2010) and detailed annotation are concerned with quality assurance and differentiating between morpho-syntactic inflection, modification, and phonological variation as a basis for the lexicographic analysis and description of signs.

In the following we will concentrate on the *Public DGS Corpus*. For information on the development of the dictionary DGS–German see Langer et al. (2018) and Wähl et al. (2018). For a discussion on how to link corpus and dictionary see Müller et al. (2020).

2.1. The Public DGS Corpus: One Corpus, Two Portals

The *Public DGS Corpus* is a 50 hour subset of the *DGS Corpus* intended for public release. In order to address the different needs of varying user groups (Jahn et al., 2018) access is provided via two different portals. As the DGS-Korpus project follows an open-access policy, both portals are freely accessible without any registration.

The first portal, *MY DGS*, addresses those interested in DGS, the history, life and culture of the deaf community. It contains over 47 hours of videos selected from the core elicitation tasks "Free conversation", "Discussion", "Subject areas", "Experience reports", "Region of origin", and "Deaf events" with German translations as optional subtitles, plus 2.4 hours of jokes (without translation).

The other portal, *MY DGS – annotated*, aims at an international audience that is interested in DGS data to perform their own research. In addition to the recordings of *MY DGS* it also contains 1.7 hours of recordings covering the remaining research-oriented elicitation tasks. These are included to provide examples of the variety of tasks in the *DGS Corpus*. Only two tasks are not part of the *Public DGS Corpus*: "Sign names" (for reasons of anonymisation) and "Isolated items" (elicitation using word and/or picture prompts). The videos are annotated with lemmas, mouthings/mouth gestures and translations. The research portal makes the videos and their annotations accessible both through a variety of downloadable file formats and through the portal website itself (see Section 4.1).

2.2. Release history

Videos and annotations of the *Public DGS Corpus* are being released and extended progressively throughout the life time of the project. To begin with, a pre-release of *MY DGS* containing ten hours of recordings was published in December 2015. Throughout 2016 and 2017 further recordings were added and improvements to the website were implemented. In May 2018 the first full release of the *Public DGS Corpus* was published (Jahn et al., 2018). This involved a content update to *MY DGS* and the first release of *MY DGS – annotated*. It increased the number of recording hours to 45.5. In February 2019 the annotation conventions of the corpus were added (Konrad et al., 2018).

Release 2, which was timed to coincide with with the TISLR 13 conference in September 2019, reached the project's target goal of 50 hours of publicly accessible recordings with almost 49 hours of lemmatised videos and more than 373,800 tokens. Starting with this release and continuing with Release 3, the focus of the *Public DGS Corpus* was shifted from adding size to adding depth.

3. New Features

In the following we report changes introduced in Release 2 and 3, focusing on new features. In Section 4 we also discuss the data formats of the corpus, the selection of which has also grown over time.

3.1. Persistent Identifiers and Archival Copies

As new versions of the *Public DGS Corpus* are released, previous versions are moved to publicly accessible archive directories. This raises the challenge of allowing users to clearly and persistently identify the version of the resource, e. g. for the purpose of citation. If a scientific article were to cite the research portal only by its URL, readers following the link could not be certain that they were viewing the same version of the corpus that the research of the article was based on, which would in turn affect the reproducibility of the research.

Instead, it has become good practice to identify resources by a persistent identifier, such as a **digital object identifier (DOI)**. DOIs allow objects, such as specific versions of a dataset, to be uniquely identified. A given DOI should always point to one and the same object and each object should only ever have one DOI. Different versions of an object should have different DOIs. A DOI is bound to metadata about its object, such as a URL at which it can be found. When the URL of the object changes, the metadata is updated to reflect this change, while the DOI stays the same. The metadata of a DOI can also be used to provide a description of the resource, citation information, version information and to connect it to the DOIs of other versions of the same data or to related resources.

In Release 2 we introduce DOIs for each release of *MY DGS* and *MY DGS – annotated*. Apart from DOIs for the overall web portals, we also provide DOIs for each individual video on the community portal and for each transcript and each sign type on the research portal. As part of this step, DOIs were generated not only for new releases, but also for the original first release.

Of course, there are also cases in which one might wish to refer to a resource in general, rather than to a specific version, e. g. to refer to a video on *MY DGS* via DOI (to be protected from changing URLs) while at the same time profiting from possible future corrections to its subtitles. For such purposes, Release 3 introduces **Concept DOIs**, which are DOIs that always point to the latest release of an object. Concept DOIs are created for all objects for which version-specific DOIs were previously created.

An important remark regarding our understanding of persistence: The corpus releases provide **semantic persistence**, but do not guarantee **byte persistence**. Semantic persistence refers to the semantic information that a resource provides. In the case of the corpus portals, this covers information like its recordings, their transcripts, the type names and token-type structure, keyword index, the reported statistics on data collection regions and informants

or the cross-references to other resources introduced in this paper (see Section 3.4). Any change to these kinds of information, regardless of whether it is the addition of new recordings or the correction of a single annotation or translation, results in a new release and the archival of the previous version.

Byte persistence, on the other hand, also implies that every byte of the digital files of a resource remains unchanged. This is undesirable for our purposes. For example, the HTML code of a page may have to be changed to ensure that it is rendered on modern web browsers or to update the hyperlink to another resource. These changes do not affect the semantic content of our resource and refraining from applying them would eventually result in archived releases becoming unusable due to purely technical reasons.

There are also certain components of the portal websites that we do not consider to be part of the semantic content of our resource for the purpose of persistence. These include, for example, the legal information provided in the imprint and data privacy information, which might have to be updated to comply with legal requirements. We also expressly exclude the title page of the community portal, as it presents a regularly changing selection of content, such as seasonal greetings and topic-specific compilations of stories from the corpus, e. g. how deaf Germans experienced the fall of the Berlin Wall.

It should also be noted that the exact specifications for data persistence of the *Public DGS Corpus* were determined during the months following Release 1 and a few additions were applied to the research portal retroactively. These changes, such as the linking of the annotation conventions (Section 3.3) and adding a second interface language (Section 3.7) are presented in this paper. None of these changes affect the corpus data itself. The archival version of Release 1 therefore represents the state of the research portal in February 2019.

3.2. Pose Information

To allow the computational processing of signed dialogues, we provide explicit machine-readable information on the location of various body parts, such as hands, shoulders, nose, ears, individual finger joints etc. This information is generated automatically using the pose estimation tool OpenPose (Cao et al., 2019). Apart from a general body model, which identifies major keypoints, such as elbows, shoulders, wrists, hip joints, eyes, nose, or ears, OpenPose can also compute detailed models of the face and each hand (Simon et al., 2017). An example of the computed information can be seen in Figure 1.

In Release 2 we introduced 2-dimensional pose information for perspectives *A* and *B*, the two frontal recordings of the participants. Release 3 adds pose information for perspective *Total*, which shows a side-view of both participants as they face each other. While the perspective also shows the moderator sitting between the participants and facing the camera, we choose not to include them in the pose information, as the moderator is not part of the corpus annotation (apart from translations of moderator utterances to aid the general understanding of the flow of conversation).

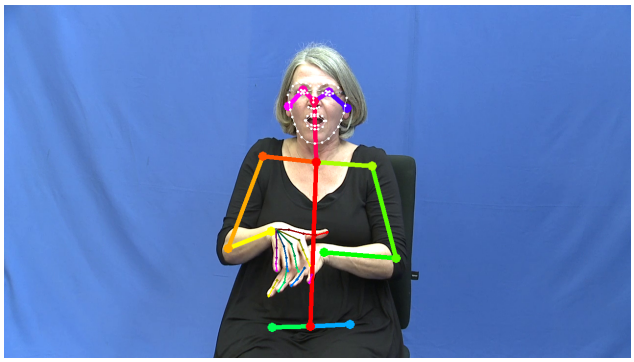

Figure 1: Visual representation of the pose information provided by OpenPose, computed for a video from the DGS-Korpus project. Sets of keypoints are generated for the body, the face and each hand. Lines between the points are added to the visual representation to indicate the logical connection between individual keypoints.

3.3. Annotation Conventions

The annotation conventions[5] were originally published as part of Release 1, explaining our approach of using a type hierarchy (double glossing) and double-token tags in iLex and the glossing conventions in *MY DGS – annotated* (Konrad et al., 2018). They were updated for Release 2 to report a change to the type-subtype relation of lexicalised forms of signs based on a manual alphabet, initialisation, and cued speech. In Release 3 we added the concordance view of tokens in each type entry (see Section 3.5) and introduced a *Sign/Lexeme* tier for each hand. Consequently, the description of "double-token tags" was updated.

3.4. Cross-References

The types list of *MY DGS – annotated* is automatically generated and shows all types and subtypes of the public corpus. For each type and its subtypes all tokens are listed below their respective gloss. In case that a studio reproduction of the citation form of the sign is available, the video is displayed under the gloss name. Studio recordings made for the *DW-DGS* show the isolated sign in four perspectives. Videos from prior productions provide a single perspective. As more dictionary entries are produced for the *DW-DGS*, more videos will also be added to the type list.

Release 3 also adds cross-references to lexical resources, namely to the *DW-DGS* and the language-for-specific-purposes (LSP) dictionaries *GaLex* (Konrad et al., 2010), *GLex* (Konrad et al., 2007), and *SLex* (Hanke et al., 2003).[6] Apart from the general value of cross-referencing resources, this also helps to contrast the difference between the type entries of *MY DGS – annotated* and the full lexical entries of the *DW-DGS*. For an example of a type entry with a multi-perspective video and cross-references, see the entry of `SMOOTH-OR-SLICK1^`[7] in Figure 2 . For a more in-depth discussion of our cross-referencing efforts, see Müller et al. (2020).

[5] `https://doi.org/10.25592/uhhfdm.822`

[6] Note that these lexical resources are not available in English.

[7] `https://doi.org/10.25592/dgs.corpus-3.0-type-13082`

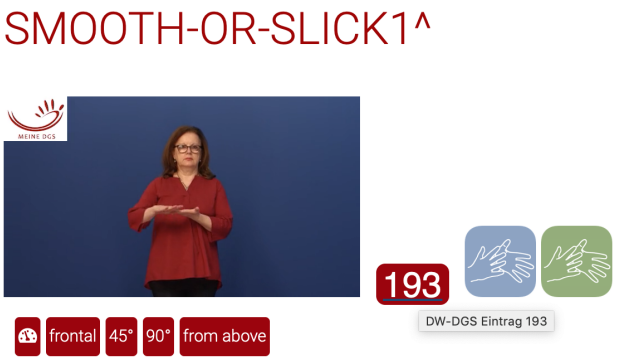

Figure 2: View of the top part of the type entry for SMOOTH-OR-SLICK1^ on the research portal website. The video image shows the citation form of the sign. Below it are buttons to change the video perspective. To the right of the video are cross references to lexical entry 193 of the *DW-DGS* and to entries in *GLex* and *GaLex*.

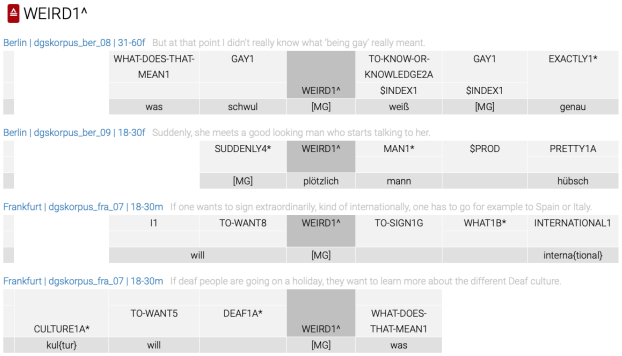

Figure 3: Concordance view for type WEIRD1^, showing the first four tokens of the type in their context.

3.5. KWIC Concordance

Type entries on *MY DGS – annotated* list tokens for each type under the type or subtype gloss. In Releases 1 and 2, each token was represented as its gloss name and relevant metadata (region, format, age group, and sex) like e. g. EMBARRASSING2 Stuttgart | dgskorpus_stu_13 | 31-45f (Release 2 removes the indication of the elicitation task that was included in Release 1 (cf. Jahn et al., 2018)). The metadata information is also a hyperlink to the occurrence of the token within the transcript.

Release 3 significantly extends the type entries by displaying each token in a keyword-in-context (KWIC) concordance. KWIC concordance is a well-known tool in corpus linguistics in which a list of tokens of the search item (e. g. a word form or any annotated information) is given with its immediate context (items before and after, i. e. left and right neighbours). The list is centred around the search item. In the case of our type entries the KWIC concordance displays the tokens of each type and subtype with up to three neighbours left and right of the searched item. The metadata is displayed above the KWIC concordance. Next to it the translation of the utterance to which the token belongs is provided to give additional context. The gloss name of the token is integrated directly into the KWIC concordance.

Figure 3 shows the concordance of the first four tokens listed for the type WEIRD1^[8]. The entry for the first token is headed by its metadata, Berlin | dgskorpus_ber_08 | 31-60f, and the English translation of the utterance it is part of, "But at that point I didn't really know what 'being gay' really meant.". The translation tag limits the range out of which the left and right neighbours of the target token are taken. That's why some concordances show less than three neighbour tokens left or right.

In addition, the KWIC concordance specifies the hand(s) the signer uses. The uppermost row displays a gloss when the sign is executed with the right hand, the middle row is for the left hand. The bottom row shows simultaneously articulated mouthings or mouth gestures. In the case of a mouthing spreading across multiple tokens, the annotation is centred in relation to the respective tokens.

For two-handed signs, the left and right hand rows are merged. As can be seen in Figure 3, when the dominance in two-handed symmetrical signs can be identified, the gloss is aligned with the row representing the dominant hand. Otherwise it is centred between the two rows. Consequently, the online transcript now also contains separate "Lexeme/Sign" columns for each hand (see Section 4.1.1).

KWIC concordances usually have a built-in function to sort the lines by left or right neighbour in order to look for collocations. This sort function is also be implemented in Release 3. Except for the target type, glosses in the KWIC concordance are clickable – like in the online transcript view – to open the respective type entry of the types list. The concordance view is also implemented in the dictionary entries of the *DW-DGS* (cf. Müller et al., 2020).

3.6. Usability on mobile devices

Since their introduction, smartphones and tablets have become more and more popular, replacing traditional desktop computers in many areas of life. In 2016 the internet use via mobile device passed desktop use for the first time (Statcounter GlobalStats, 2016). Making websites mobile-friendly has become a priority for web design. This requires layouts compatible with various screen sizes and touch-compatible navigation.

Furthermore, advances in web standards, such as the introduction of HTML 5, enabled the design of websites that are almost indistinguishable from regular mobile applications. Modern mobile operating systems even allow users to store websites as de-facto apps, representing them as a dedicated app icon on the home screen and hiding the web browser interface.

The *Public DGS Corpus* portals have always been designed with mobile-friendliness (and accessibility) in mind. As of Release 3, both portals can now also be stored as mobile apps, making access to the corpus even more seamless.

[8] https://doi.org/10.25592/dgs.corpus-3.0-type-18560

3.7. Update to Interface Languages

While the *Public DGS Corpus* was published with annotations in both German and English, the user interfaces of its portals were originally monolingual. The community portal was provided in German, as it is the written language most accessible to native speakers of DGS. The research portal had an English user interface, the lingua franca of the international research community. The only exception to this was the transcript web viewer (see Section 4.1.1), which allowed users to switch between German and English transcriptions (Jahn et al., 2018).

However, a side-effect of these language choices was the unintended implication that community members should not also be interested in the linguistic information available only on the research portal. Thanks to user feedback we became aware of this issue and upgraded the research portal user interface to be fully available in both English and German. This upgrade was also retroactively applied to Release 1, as it did not change the corpus itself and was deemed an urgent correction that should not wait until Release 2.

4. Data Formats

The page *Transcripts* on the research portal provides a list of all the dialogue transcripts that are part of the *Public DGS Corpus*. For each transcript it provides metadata (age group, elicitation format, topics of conversation) and a variety of file formats in which to access the corpus data. There are different file formats for annotation data (Section 4.1), video data (Section 4.2), pose information (Section 4.3) and metadata (Section 4.4).

4.1. Annotation Data

To support a variety of linguistic tools, the corpus annotations are made available in a number of different formats. They can be accessed via a web viewer (Section 4.1.1) or downloaded as *iLex*, *ELAN* or *SRT* files (Sections 4.1.2 to 4.1.4, respectively). All formats contain the basic annotation, i. e. translations, type glosses, and mouthings/mouth gestures. The inclusion of additional information depends on the limitations of each format.

4.1.1. Web Viewer

The web viewer provides a fast and easy way to directly inspect the *Public DGS Corpus* data without having to download it. It is is reachable via the research portal *MY DGS – annotated*, by clicking on the name of a transcript. It was first described by Jahn et al. (2018).

In the web viewer, the two informants are presented side by side in a video at the top of the page. Beneath it, the transcript is shown in vertical form (time flowing from top to bottom). The header of the transcript provides its name and a list of covered topics. In the vertical transcript, three tiers per informant are displayed: a translation (in either German or English), the tier *Lexeme/Sign* showing the glosses (in German or English) and the *Mouth* tier that displays mouthings (in German) or "[MG]" for any mouth gestures of the informant. A last column displays utterances or actions of the moderator as far as they are of (potential) relevance for the conversation.

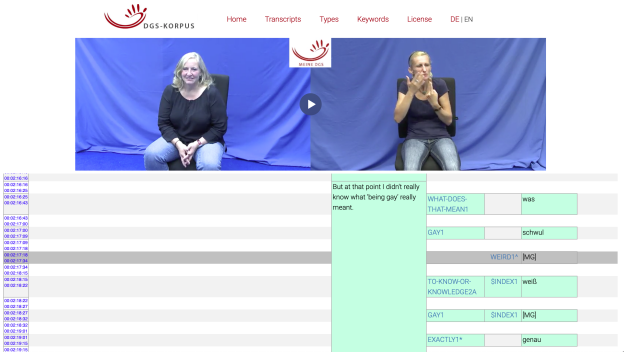

Figure 4: Web view of a transcript, with the video paused at a token of `WEIRD1^` articulated with the left hand being the active one.

Release 2 added DOIs for the respective transcript, placed above the transcript name. In Release 3, the *Translation* and *Mouth* tiers remain unchanged, but the *Sign/Lexeme* tier is now presented in two columns, one for the right and one for the left hand, as can be seen in Figure 4. This ensures consistency between the web viewer and the KWIC concordance view described in Section 3.5.

4.1.2. iLex

iLex (Hanke, 2002; Hanke and Storz, 2008) is an integrated annotation tool and lexical database, specifically designed to support consistent token-type matching (lemmatisation) and further annotation of sign language texts. The *DGS Corpus* and *Public DGS Corpus* were created using iLex, so it is naturally the tool that can model their information most accurately. Individual tokens refer to underlying type entries in the lexical database that are hierarchically structured into types and subtypes (Konrad et al., 2012; Konrad et al., 2018). The iLex files are the only available format that can explicitly represent the token-type relation and the type-subtype hierarchy.

Apart from a gloss, types also contain a phonetic transcription of their citation form using HamNoSys (Hanke, 2004). This should not be confused with a token transcription, which would also take into account deviations from the citation form. As iLex is the only one of our formats that can explicitly model the difference between tokens and types, and therefore the difference between token transcriptions and type transcriptions, we provide phonetic transcriptions only in this format.

Each transcript of the *Public DGS Corpus* is made available as an iLex *XML* file. Each file contains translations, gloss tokens, mouthings/mouth gesture annotations, gloss type hierarchies and phonetic HamNoSys transcriptions of types. When working with iLex the video recordings associated with an annotation can either be stored locally or accessed remotely on the *Public DGS Corpus* server.

4.1.3. ELAN

ELAN[9] (Crasborn and Sloetjes, 2008) is another popular annotation tool for sign language annotation. Information in ELAN is represented in tiers which are time-aligned to video files. The first time an .eaf file downloaded from *MY*

[9] `https://tla.mpi.nl/tools/tla-tools/elan/`

DGS – annotated is opened with ELAN, the location of the video files for the *A*, *B* and *Total* perspectives must be set. If files are not available locally, the users can also choose to work with fewer or none of the videos.
Each *ELAN* file provides 24 tiers that contain translations, lemmatisation and mouthings/mouth gestures. For each kind of information there exists a tier in English and in German. The only exception is the *Mouthing/Mouth Gesture* tier. Mouthings were not translated in English, as they refer to German words with different articulation features from e. g. mouthed English words. Within one language, one *Translation* tier each is provided for informant A, informant B and the moderator.
The lemmatisation by means of glosses is displayed in four tiers per informant and language. These result from the type hierarchy in iLex (two tiers for double glossing; see Konrad et al. (2018)) and the distribution to the active hand (two tiers for left and right hand):

Type hierarchy: In ELAN, type hierarchies are supported only indirectly. As ELAN does not support explicit type relations, each token is represented as the gloss of its type in a *Sign* tier. If the token has a subtype then the subtype gloss is represented in a *Lexeme/Sign* tier. For tokens connected directly to a type, the *Lexeme/Sign* tier is left empty.

Active hand: Depending on which hand is active in articulating the sign, tokens appear in tiers for either the right or left hand. For two-handed asymmetric signs the tiers of the active hand are filled. In the case of symmetric signs, the dominance of the hand is determined where possible, otherwise the right hand tiers are filled as default. Also, glosses for nonmanual activity like nonmanual gestures or exclusively oral activity are displayed in the right hand tiers (Konrad et al., 2018).

In ELAN, glosses represent tokens. The hierarchical relation between types and subtypes is lost. Therefore, information applied only to types but not tokens, such as the HamNoSys notation of citation forms, is not included in this format.

4.1.4. SRT

The SubRip Subtitle file format (*SRT*) is a popular format for storing subtitles separately from their video file. We provide our core annotation in this general-purpose format to allow its use with additional tools, such as MaxQDA[10], and in regular media players. In *SRT* files, text strings are associated with start and end timestamps to determine the time span in the video during which they should be displayed. It does not permit the inclusion of meta-information or the inclusion of multiple tracks to differentiate information. This means that there is no technical difference between type (glosses), mouthing, and translation items. To at least identify the origin of each utterance, each subtitle element starts with the identifying letter of its speaker (*A* and *B* for the participants, *C* for the moderator). The German and English data are provided in separate files.

[10] https://www.maxqda.com/

4.2. Video Data

All corpus recordings are provided as *MP4* video files, encoded using *H.264* compression at a resolution of 640 by 360 pixels and 50 frames per second.
Three perspectives are available: *Video A* and *Video B* each provide a frontal view of participant A and B, respectively. *Video Total* shows both participants from their side, facing each other, with the moderator sitting between them, facing the camera.
A fourth file, called *Video AB*, shows perspectives *A* and *B* next to each other. It corresponds to the video format shown in the web viewer (see Section 4.1.1) and on *MY DGS*. This file is provided for users that use the *SRT* format in applications that can only play a single video at a time.

4.3. Pose Information

The pose information for each transcript (see Section 3.2) is provided as a *JSON* file. To reduce its file size during transfer, the file is compressed using *gzip*. While OpenPose by default generates individual files for each frame, we compile all frames of all video perspectives in a single file. For users who require the default one-file-per-frame format, we provide a conversion script.[11]
Apart from the OpenPose output, the file also includes relevant metadata, such as the transcript ID, the camera perspective and the pixel dimensions of the original video on which OpenPose was run. Pixel dimensions are particularly important for users who wish to apply the pose information to the video files found on the research portal (see Section 4.2), as these are of smaller resolution than the original videos.
For further details on the OpenPose data of the *Public DGS Corpus* and its file format, please see the project note by Schulder and Hanke (2019).[12]

4.4. Metadata

Any kind of language resource will naturally have various kinds of metadata associated with it. This can be resource-wide information, like which language or languages the corpus contains, or information on specific parts, such as which age group individual informants belong to. To provide a standard for describing such language resource metadata, the Component MetaData Infrastructure (*CMDI*) was introduced in ISO 24622-1:2015 (2015).
The data formats we provide for the corpus have varying degrees of support for including metadata. To provide a single independent source for metadata, Release 3 introduces *CMDI XML* files for every transcript.

5. Outlook

One of the main motivations for many decisions regarding the design of the *Public DGS Corpus* and the changes made in its release versions was the feedback of users of the *Public DGS Corpus*. This shows for example in the changes the web viewer underwent throughout the releases or the addition of German as an interface language for *MY*

[11] https://github.com/DGS-Korpus/Public-Corpus-OpenPose-frame-extractor
[12] https://doi.org/10.25592/uhhfdm.842

DGS – annotated to not exclude user groups without English skills. While the web viewer of *MY DGS – annotated* was intended as a preview of the data that helps researchers select suitable data for download and further analysis with tools like iLex or ELAN, it turned out that many users prefer using the web viewer and expect to be able to do their research in it directly.

We look forward to feedback on new features such as the KWIC view. While we expect these features to mature over time and become sufficient for many purposes, this by no means replaces a full corpus research tool. As was announced in Jahn et al. (2018) we are also working on providing our data for ANNIS[13] (ANNotation of Information Structure).

6. Acknowledgements

We would like to thank the many student annotators who helped create the corpus and who contributed many of the corrections in its subsequent releases.

Also, we are very thankful for the valuable feedback provided by the community.

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 Academies of Sciences and Humanities.

7. Bibliographical References

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv preprint, v2.

Crasborn, O. and Sloetjes, H. (2008). Enhanced ELAN Functionality for Sign Language Corpora. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 39–43, Marrakech, Morocco. European Language Resources Association.

Hanke, T. and Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 64–67, Marrakech, Morocco. European Language Resources Association.

Hanke, T., Storz, J., and Wagner, S. (2010). iLex: Handling Multi-Camera Recordings. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 110–111, Valletta, Malta. European Language Resources Association.

Hanke, T. (2002). iLex – A tool for Sign Language Lexicography and Corpus Analysis. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 923–926, Las Palmas, Canary Islands, Spain. European Language Resources Association.

Hanke, T. (2004). Hamnosys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 1–6, Lisbon, Portugal. European Language Resources Association.

ISO 24622-1:2015. (2015). Language resource management — Component Metadata Infrastructure (CMDI) — Part 1: The Component Metadata Model. Standard, International Organization for Standardization, Geneva, Switzerland.

Jahn, E., Konrad, R., Langer, G., Wagner, S., and Hanke, T. (2018). Publishing DGS Corpus Data: Different Formats for Different Needs. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 107–114, Miyazaki, Japan. European Language Resources Association.

König, S., Konrad, R., Langer, G., and Nishio, R. (2010). How Much Top-Down and Bottom-Up Do We Need to Build a Lemmatized Corpus? *Poster presented at the International Conference on Theoretical Issues in Sign Language Research*, West Lafayette, Indiana, USA.

Konrad, R. and Langer, G. (2009). Synergies between Transcription and Lexical Database Building: The Case of German Sign Language (DGS). In *Proceedings of the Corpus Linguistics Conference*, Liverpool, United Kingdom. University of Liverpool.

Konrad, R., Hanke, T., König, S., Langer, G., Matthes, S., Nishio, R., and Regen, A. (2012). From Form to Function. A Database Approach to Handle Lexicon Building and Spotting Token Forms in Sign Languages. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 87–94, Istanbul, Turkey. European Language Resources Association.

Konrad, R., Hanke, T., Langer, G., König, S., König, L., Nishio, R., and Regen, A. (2018). Public DGS Corpus: Annotation Conventions. Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Langer, G., Müller, A., Wähl, S., and Bleicken, J. (2018). Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 483–497, Ljubljana, Slovenia. Ljubljana University Press.

Müller, A., Hanke, T., Konrad, R., Langer, G., and Wähl, S. (2020). From Dictionary to Corpus and Back Again – Linking Heterogeneous Language Resources for DGS. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, Marseille, France. European Language Resources Association.

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T., and Rathmann, C. (2010). Elicitation Methods in the DGS (German Sign Language) Corpus Project. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 178–185, Valletta, Malta. European Language Resources Association.

Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., and Schwarz, A. (2008). DGS Corpus Project – Development of a Corpus Based Electronic Dictionary German Sign Language / German. In *Proceedings of the Work-*

[13] http://corpus-tools.org/annis/

shop on the Representation and Processing of Sign Languages at LREC, pages 159–164, Marrakech, Morocco. European Language Resources Association.

Schulder, M. and Hanke, T. (2019). OpenPose in the Public DGS Corpus. Project Note AP06-2019-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand Keypoint Detection in Single Images Using Multiview Bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653.

Statcounter GlobalStats. (2016). Mobile and tablet internet usage exceeds desktop for first time worldwide. https://gs.statcounter.com/press/mobile-and-tablet-internet-usage-exceeds-desktop-for-first-time-worldwide, November. Accessed: 2020-04-01.

Wähl, S., Langer, G., and Müller, A. (2018). Hand in Hand – Using Data from an Online Survey System to Support Lexicographic Work. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 7–12, Miyazaki, Japan. European Language Resources Association.

8. Language Resource References

Hanke, T., Konrad, R., Schwarz, A., König, S., Langer, G., Pflugfelder, C., and Prillwitz, S. (2003). *Fachgebärdenlexikon Sozialarbeit/Sozialpädagogik*. Arbeitsgruppe Fachgebärdenlexika, IDGS, Hamburg University, URL http://www.sign-lang.uni-hamburg.de/slex/.

Hanke, T., König, S., Konrad, R., Langer, G., Barbeito Rey-Geißler, P., Blanck, D., Goldschmidt, S., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., König, L., Matthes, S., Nishio, R., Rathmann, C., Salden, U., Wagner, S., and Worseck, S. (2018). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 1. Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.meinedgs-1.0.

Hanke, T., König, S., Konrad, R., Langer, G., Barbeito Rey-Geißler, P., Blanck, D., Goldschmidt, S., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., König, L., Matthes, S., Nishio, R., Rathmann, C., Salden, U., Wagner, S., and Worseck, S. (2019). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 2. Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.meinedgs-2.0.

Hanke, T., König, S., Konrad, R., Langer, G., Barbeito Rey-Geißler, P., Blanck, D., Goldschmidt, S., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., König, L., Matthes, S., Nishio, R., Rathmann, C., Salden, U., Wagner, S., and Worseck, S. (2020). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.meinedgs-3.0.

Konrad, R., Langer, G., König, S., Hanke, T., and Prillwitz, S. (2007). *Fachgebärdenlexikon Gesundheit und Pflege*. Arbeitsgruppe Fachgebärdenlexika, IDGS, Hamburg University, URL http://www.sign-lang.uni-hamburg.de/glex/.

Konrad, R., Langer, G., König, S., Hanke, T., and Rathmann, C. (2010). *Fachgebärdenlexikon Gärtnerei und Landschaftsbau*. Arbeitsgruppe Fachgebärdenlexika, IDGS, Hamburg University, URL http://www.sign-lang.uni-hamburg.de/galex/.

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., and Worseck, S. (2018). *MY DGS – Annotated. Public Corpus of German Sign Language, 1st Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.corpus-1.0.

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., and Worseck, S. (2019). *MY DGS – Annotated. Public Corpus of German Sign Language, 2nd Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.corpus-2.0.

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., and Schulder, M. (2020). *MY DGS – Annotated. Public Corpus of German Sign Language, 3rd Release*. DGS-Korpus project, IDGS, Hamburg University, DOI 10.25592/dgs.corpus-3.0.