Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages, pages 157–164
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

From Dictionary to Corpus and Back Again – Linking Heterogeneous Language Resources for DGS

Anke Müller, Thomas Hanke, Reiner Konrad, Gabriele Langer, Sabrina Wähl
Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Germany
{anke.mueller, thomas.hanke, reiner.konrad, gabriele.langer, sabrina.waehl}@uni-hamburg.de

Abstract

The *Public DGS Corpus* is published in two different formats, that is subtitled videos for lay persons and lemmatized and annotated transcripts and videos for experts. In addition, a draft version with the first set of preliminary entries of the DGS dictionary (*DW-DGS*) to be completed in 2023 is now online. The *Public DGS Corpus* and the *DW-DGS* are conceived of as stand-alone products, but are nevertheless closely interconnected to offer additional and complementary informative functions. In this paper we focus on linking the published products in order to provide users access to corpus and corpus-based dictionary in various, interrelated ways. We discuss which links are thought to be useful and what challenges the linking of the products poses. In addition we address the inclusion of links to other, older lexical resources (LSP dictionaries).

Keywords: dictionary, corpus, cross-linking

1. Introduction

The DGS-Korpus project is a long-term project (2009-2023) that has three major aims: a) compiling a reference corpus of German Sign Language (DGS), b) publishing part of the annotated corpus, c) compiling and publishing a corpus-based dictionary DGS–German. Data collection took place from 2010 to 2012 and captured near-natural DGS data from 330 informants coming from all over Germany (Nishio et al., 2010). The *DGS Corpus* contains about 560 hours of DGS signing. Lemmatising and annotating is done with `iLex`[1] (Hanke, 2002; Hanke and Storz, 2008), a lexical database and annotation tool designed for a multi-user environment. A subset of about 50 hours was selected for publication. This *Public DGS Corpus* was published on two different portals, *MY DGS*[2] and *MY DGS – annotated*[3]. The corpus-based dictionary *Digitales Wörterbuch der Deutschen Gebärdensprache* (*DW-DGS*) is still in the making. Its final version is to be published end of 2023. In order to test and discuss form, content, and usability with the language and the research community, we make a pre-release of dictionary entries available[4]. Since the *DW-DGS* and the *Public DGS Corpus* are closely related, it is obvious to make the relation tangible for the users of both *DW-DGS* and *Public DGS Corpus*. In addition, we want to integrate information on DGS signs that was published earlier in several LSP (language for specific purposes) dictionaries German–DGS. Thus, several features link dictionary, corpus, and heterogeneous DGS language resources.

2. Data Structure and Language Resources

2.1. Data Structure

In iLex, types are database entities with unique IDs, which tokens are linked to. A type is an abstract unit of the language with a specific form that – for iconically motivated signs – is associated with a specific underlying image (König et al., 2008). Its form can have several realisations in actual use and it can have a number of different conventional meanings. In order to group tokens according to these conventional meanings, we implemented a type hierarchy (type levels) and double glossing: Each type (parent) can have one or several subtypes (children) (Konrad et al., 2018; Langer et al., 2016). At the beginning of the lemmatisation of the *DGS Corpus* data two additional type levels – qualified types and qualified subtypes (Konrad et al., 2012) – were implemented to group recurrent form variations and modifications of types or subtypes. Tokens are matched either to a type, a subtype or a qualified type. A type entity in iLex is defined at least by a gloss and a citation form in HamNoSys[5] (Hanke, 2004). Type and subtype glosses are given in *MY DGS – annotated*, whereas qualified types are used but internally in the *DGS Corpus*.

When the DGS-Korpus project started, iLex already comprised a large number of type entities and lemmatised tokens of collected data as well as of studio reproductions of isolated signs (citation form). For the production of LSP dictionaries (see Section 2.4) quite an amount of supplementary production data were available.

As before, the *Public DGS Corpus* (Section 2.2) is produced from the data stored, managed and prepared in iLex. This also applies for the *DW-DGS*. The data includes types selected for dictionary entries, studio reproductions for representing the signs' citation forms, and video sequences taken from the *DGS Corpus* to serve as examples for sign senses described in the respective entry (Langer et al., 2018).

One of the first steps when compiling a dictionary is to define which data from the corpus is to be covered by and described in a dictionary entry, that is, which types or parts of a type structure should be included. This step is called

[1] `https://www.sign-lang.uni-hamburg.de/ilex/`
[2] `http://meine-dgs.de`
[3] `http://ling.meine-dgs.de`
[4] `http://dw-dgs.meine-dgs.de`
[5] `http://www.dgs-korpus.de/index.php/hamnosys-97.html`

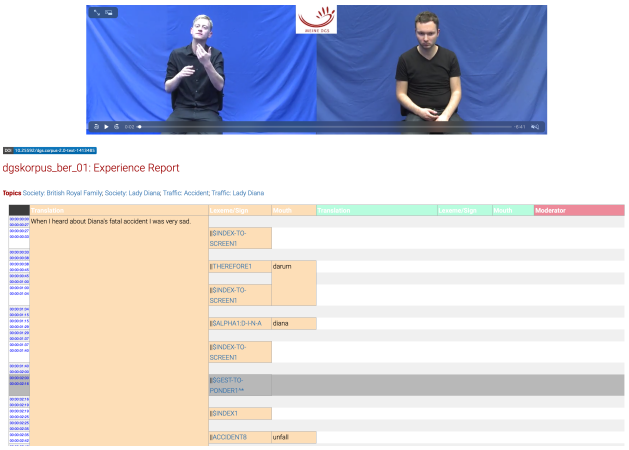

Figure 1: Online transcript view. Location: Berlin. Format: Experience Report. Topics: Society – Lady Diana.

lemma establishment (Svensén, 2009, p. 94). Rules for lemma establishment and annotation guidelines (for token-type matching, i. e. lemmatisation) serve different purposes, may follow different rules and reflect different stages of analysis (Langer et al., 2016). Lemma establishment decisions reflected in the scope of dictionary entries may not necessarily lead to changes in the data structure of type entities in iLex. Thus, a dictionary entry can cover more than one type or a type in iLex can be split up into more than one dictionary entry. It is also possible that a branch of a type structure in iLex is described in a separate entry together with data from another type. While this is unproblematic for the dictionary as a stand-alone product, it makes interlinking of corpus and dictionary more challenging.

2.2. The Public DGS Corpus

The prerequisite for building the *DGS Corpus* was the consent of the informants to collect, analyse and publish their data. In order to give something back to the language community and to make data accessible for the sign language research community, one project goal is the publication of about 50 hours of signing with annotations in German and English. We decided to publish this *Public DGS Corpus* in two portals suited for the different needs of the language and the research community. Selection, processing steps, data formats and features of Release 1 are reported in Jahn et al. (2018), for the changes in Releases 2 and 3 see Hanke et al. (2020). With Release 2, the target quantity of 50 hours was reached.

2.2.1. Portal *MY DGS*

This portal addresses users who are interested in the content of discussions, conversations, and narratives on history, life and culture of Deaf people. It contains over 47 hours of videos with translations from DGS to German as optional subtitles. *MY DGS* provides a low-threshold access to the data. The metalanguage used for description on this website is German only. In addition, 2.4 hours of jokes as part of the (German) deaf culture can be browsed (without subtitles). The videos can be filtered for 13 regions, 4 age groups, 8 formats (elicitation tasks) and 38 main topics. For the following, the 47 hours of discussions and conversations are of interest as there are links from the *DW-DGS* examples to these sequences (see Section 3.2).

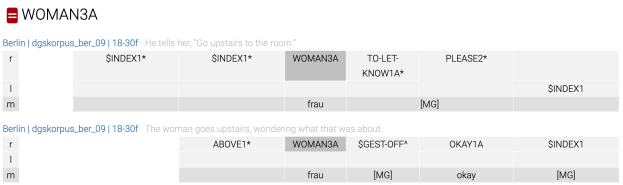

Figure 2: KWIC concordance with tokens of the subtype WOMAN3A.

2.2.2. Portal *MY DGS – annotated*

This portal made for the research community includes the video material of *MY DGS* lemmatised and annotated (except jokes) and is fully available in both English (except mouthings) and German. Of the tasks not included in *MY DGS* additional 1.7 hours of video were selected to exemplify the whole range of tasks covered in the data collection. We considered this material more important for research than for the general public. Only the tasks "Sign Names" and "Isolated Items" are not part of the *Public DGS Corpus*. In the following, we focus on the online transcript view and the types list of *MY DGS – annotated*, for detailed information on data formats and features see Hanke et al. (2020).
The online transcript shows a video with both informants in the frontal camera perspective (during the elicitation they were sitting face to face). Beneath, versioning (DOI), video name (with location/region and task) and topics are given, followed by a vertical transcript as shown in Figure 1.

All glosses in the "Lexeme/Sign" tier are clickable leading to the corresponding type or subtype section of the type entry. Figure 2 shows the section of the subtype WOMAN3A where all tokens of this subtype in the *Public DGS Corpus* are listed as a KWIC concordance (keyword in context) and highlighted by a dark grey shade: Each token is listed by its metadata and the English translation of the utterance it is part of, that is for the first token of WOMAN3A: Berlin | dgskorpus_ber_09 | 18-30f, "He tells her, 'Go upstairs to the room.'" The translation tag limits the range out of which the left and right neighbours of the target token are taken. That is why some key tokens show less than three neighbour tokens left or right. All glosses of the neighbouring tokens are clickable leading to the respective type entry. Below the token glosses annotated mouthings or mouth gestures are shown.
The parent type of WOMAN3A is EARRING1A^, which is listed at the head of this type entry (see Figure 3). In case that a studio reproduction of the citation form of the sign is available, the video is displayed under the gloss name. Studio reproductions made for the *DW-DGS* show the isolated sign in four perspectives. If the video is taken from prior productions, only one perspective is given. In the course of the production of dictionary entries more and more videos will be added.

2.3. Corpus-based DGS Dictionary

The *DW-DGS* is based on the total of annotated material of the *DGS Corpus* (with over 601700 tokens), which exceeds

Figure 3: Type entry start of EARRING1A^ with video (citation form) and links to *DW-DGS* entry 156 and same type in the gloss index of LSP dictionary *Health & Nursing*.

the published data (more than 373800 tokens). The *DW-DGS* aims at the description and documentation of signs as they are used in everyday signing, as represented in the corpus data. Though it serves the function of a bilingual dictionary with German translational equivalents and an index of German, the focus is on the description of DGS and its structures independent of German, as if in a monolingual dictionary.

The *DW-DGS* addresses diverse user groups including the language community and native signers as well as beginning and advanced learners, the general public as well as linguists. The pre-release is an incremental publication of entries along with a growing macro-structure as for example background information and search facilities. What is of interest for this paper is the structure of entries, the DGS index and the German index. The DGS index displays all entries that are fully edited by way of a micon (moving icon). One of the main design decisions for the dictionary was not to represent signs by glosses, but to use thumbnail videos and numbers instead, resulting in micons consisting of a posed still of a signing model plus a unique identification number. This prevents the user from mistaking gloss names for meaning or to confuse glosses with German, especially as German is the metalanguage for sign descriptions within the entry. The dismissal of glosses for the *DW-DGS* entries has the further advantage of avoiding a clash or discrepancy of glosses between dictionary and corpus which would occur whenever the lemma establishment does not match the lemmatisation of types in iLex. Figure 4 shows a sign entry as it appears when accessed via the DGS index. A sign entry is identified by the identification number and the citation form of the sign. Information given on a sign includes form variants of the sign, information on regional distribution, cross-references to signs with identical citation form (homonyms) and signs with similar citation forms. The main body consists of the description of the sign's senses based on the analysis of corpus data. Figure 4 shows the overview of 5 senses indicated by sign posts; each, when clicked, reveals a table of detailed information on a sense such as an explanation of meaning or usage, typically co-occurring mouthings, German translational equivalents, authentic examples directly taken from the corpus

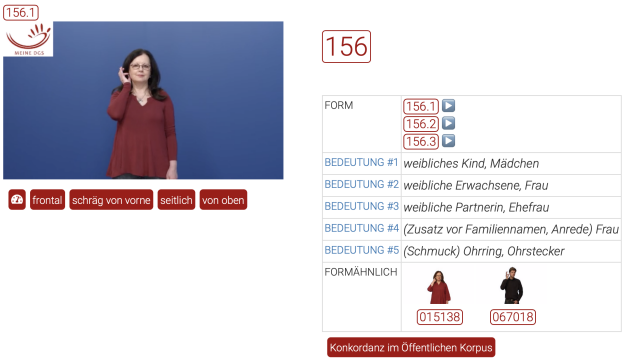

Figure 4: Entry 156 with three form variants, overview of senses with sign posts and two cross-references as micons.

for attesting and illustrating senses, cross-referenced synonyms and antonyms, and collocational patterns.

All information given in DGS can be viewed in the fixed display window, that is, the form variants of the lemma, all signs represented as micons, and examples. Micons are used for cross-references within the dictionary – when clicking the still, the corresponding film can be viewed in the film display window; the number serves as a link to the corresponding entry. A preliminary design feature is the automatic generation of entries, if there is a cross-reference to an entry that does not exist as a fully edited article. Such an automatically generated entry shows the sign form and a link back to all entries referring to the sign in question. These back links are labeled according to their relation kind, e. g. synonym of X.

The German index is a list of translational equivalents followed by entry identification numbers giving access directly to the corresponding senses indicated by the number of the sense within an entry, e. g. entry 59#2.

In the process of manually performed sense discrimination, not every token of a type is viewed and analysed, but only a critical mass to attest or confirm the most typical senses. Particularly if a sign type has many tokens, they cannot all be reviewed in detail. Moreover, not all tokens can be assigned to the senses identified, depending on the granularity of the senses. This is why, in the DGS-Korpus project, we do not have a full sense-tagging. There is no automatic solution for a reliable sense-tagging at sight. This fact has implications on the linking of dictionary and corpus (see Section 3.2).

2.4. LSP Dictionaries German–DGS

Lexicographic work on DGS was conducted at our institute previous to the DGS-Korpus project. Between 1993 and 2010, six LSP dictionaries (*Psychology*, *Joinery*, *Home Economics*, *Social Work & Social Pedagogics*, *Health & Nursing*, and *Horticulture & Landscaping*) were compiled (Konrad, 2011; Konrad and Langer, 2012). Within the context of these projects experience, methodology, know-how, and technical tools were developed and improved.

Except for the first project, DGS equivalents in the elicited answers to words and/or picture prompts and semi-structured interviews were lemmatised and annotated using annotational tools developed at our institute. The LSP dic-

tionaries are bidirectional in that they consist of two kinds of entries – concept entries with definitions headed by the German technical term, and additional sign entries of simplex signs used in the DGS equivalents of German technical terms. These signs were listed and described in sign entries accessible through sign indexes or from cross-references within the concept entries. All entries and indexes were produced directly from the information stored, corrected and prepared in a lexical database (GlossLexer Hanke et al. (2001), then iLex). In order to make the respective sign index consistent and the numbering gapless, production glosses with continuous numbering within each product partly replaced the iLex-internal glosses. As a result, glosses for the same sign may differ between the LSP dictionaries and iLex.

When the DGS-Korpus project started, iLex already comprised a large number of type entries, lemmatised tokens, annotated mouthings/mouth gestures from data collected in previous dictionary projects as well as production data and lemmatised studio reproductions of citation forms. While information on types and therefore their description in the database may have changed over time through new data, re-evaluation of data, change of annotation conventions, or corrections, there is still a considerable number of types that are used in the *DGS Corpus* data as well as in the data of previous projects. This common base of type entries can be utilised to link from entries in the types list of *MY DGS – annotated* as well as from *DW-DGS* entries to the corresponding types in the sign entries of three LSP dictionaries: *Social Work & Social Pedagogics* (Hanke et al., 2003), *Health & Nursing* (Konrad et al., 2007), and *Horticulture & Landscaping* (Konrad et al., 2010).

3. Linking Corpus and Dictionary

3.1. Challenges

Linking *MY DGS – annotated* and *DW-DGS* entails challenges that need to be considered. First, the user groups are rather diverse with different needs. The dictionary aims at a broad public interested in DGS including researchers, whereas the research portal is aimed at a scientific public. Second, as the research portal provides transcripts it also displays glosses used for lemmatisation. Within the dictionary glosses are not used to refer to signs, micons combined with numbers are used instead. These different styles may be confusing for users. Third, as Langer et al. (2016) pointed out, lemmatisation decisions in the database do not necessarily match lemma establishment in the dictionary. Hence different types from the database appearing in the *Public DGS Corpus* types list may be mapped onto one entry, or one type may be mapped onto several entries.

3.2. From Dictionary Entry to Corpus

Compiled entries of the dictionary are based on corpus occurrences. While a dictionary entry sums up forms, properties, meanings and uses of a sign, a corpus presents the data in a structured way, e.g. through a listing of all occurrences of a type and links to the source texts in annotated transcripts. The DGS-Korpus project makes both available – the results of lexicographic analysis and a structured view of tokens of the same type, which is presented as a KWIC concordance. This presentation allows users to have a look at the context a sign occurs in, as well as a comparison of left and right neighbours (for a detailed description of the KWIC concordance see Hanke et al. (2020)).

Entries in the pre-release of the *DW-DGS* contain a red button at the bottom (cf. Figure 4 or the box 'DW-DGS' in Figure 5), which when clicked opens a KWIC concordance of the tokens of all types and subtypes that constitute the respective entry, given that they occur in *MY DGS – annotated*. The view of this entry generated concordance differs from the view when accessed within *MY DGS – annotated* in some points: The list is headed by the identification number of the entry the KWIC concordance belongs to, which serves as a direct back link, and there are neither a studio reproduction nor type and subtype glosses as headings that indicate the gloss hierarchy of the iLex database (cf. box 'KWIC1' as opposed to the boxes 'KWIC2', 'KWIC3', 'KWIC4' in Figure 5). Otherwise, the same information and link structure is given with respect to the single type occurrences (tokens), that is, there is a link heading each KWIC line to the token in the respective transcript, and neighbouring glosses of the target gloss link to their respective type in another KWIC concordance (cf. arrows from KWIC1 and KWIC3). But, and this is necessarily so, the target gloss also links to the respective type in a KWIC concordance of the *MY DGS – annotated* style (KWIC3). This way a user can find out which type a particular subtype gloss may belong to.

The KWIC concordance as generated from a dictionary entry reflects the lexicographic lemma establishment, which sometimes results in sampled concordances made up from two or more types, or may also cut off a sub-branch of a type. Ideally spoken, a linguistic expert could make up their own dictionary entry by viewing all listed tokens.

Coming from the dictionary where signs are represented as stills, micons or video, the user is confronted with the use of glosses in the KWIC concordance, which they cannot directly associate to the lemma sign of the entry they may come from. If they click onto different key tokens marked by dark grey background, eventually they open all type concordances from the corpus and recognise the shown variants in the studio reproduction on top of each list, as well as the entry number of the *DW-DGS* appearing there. Though at first potentially confusing, the availability of a sampled KWIC concordance offers a lot of additional examples with a broad range of information on sign forms (modifications and phonetic variants), use and senses in different contexts, which may also include uses that are not described in the entry because they are used in a productive and sense-expanding way, or because there is too little evidence for a conventionalised use. Even the examples used in the entry may be discovered; a marking of those is a planned feature for future releases. Here, users may observe differences of segmenting and translation, which is due to our preparing an example to serve as a good example of a sense even out of context, which sometimes requires to adjust the translation of an utterance (cf. Langer et al., 2018). These adjustments are always true to the original. The examples of sign uses displayed in the KWIC concordance are not grouped according to the senses defined and listed in the

corresponding entry because tokens are not systematically sense-tagged in the corpus.
As stated above, in the pre-release of the *DW-DGS* there are many automatically generated entries without proper lemma establishment or form and sense descriptions. But they all offer the link to the corpus KWIC concordance, so a user of the dictionary can gather more information on a sign they were referred to by a cross-reference, be it a type or a subtype. Another kind of external link implemented in the dictionary entry structure is from an authentic example shown as a cut-out within the entry to the source text of the very example. Whenever an example is taken from the *Public DGS Corpus*, two red buttons show up below the video display window (see Figure 5). The first button takes the user to the beginning of the source text in *MY DGS*, where they can view the whole discourse context in full detail and observe the use of the sign of that sense in this specific case. The second button targets the beginning of the example utterance in the respective transcript of *MY DGS – annotated*.

3.3. From Corpus to Dictionary Entry

The main route leading from the *Public DGS Corpus* to the *DW-DGS* is the KWIC concordance showing all the occurrences for one type and the dependent subtypes. If there is a studio reproduction of the sign's form available, it is displayed under the gloss of the type. Next to that video you may find one or more entry numbers linking to the dictionary, if there is an entry already in existence. The number of entries linked to a type depends on lemma establishment decisions (Section 2.1) that do not necessarily map 1:1 to the type structure. Thus there are three different cases of mapping between corpus and dictionary. The simplest case is a 1:1 mapping between sign type and dictionary entry. If an entry comprises several sign types, e. g. because they are phonological variants of one another, the mapping is 1+n:1 from corpus to dictionary (see box 'KWIC2' and 'KWIC3' in Figure 5). The third case is that a subtype is defined as an entry in its own right compared to the rest of the type, e. g. because it is a sign modification with a specific meaning the other forms of the sign do not show. In that case the mapping is 1:1+n (see box 'KWIC2'). Naturally, confusion may occur especially with the third case, so information on the project's lemma establishment principles are needed in order to make the decisions transparent. The benefit for the users is that they may find information on a sign's possible meanings and uses that are not provided via the types list and concordance view directly. The dictionary also features prepared information on e. g. collocations of the sign.

4. Linking to Heterogeneous Resources

4.1. Challenges

The *Public DGS Corpus* and the *DW-DGS* are complementary products that are both based on the same data collected and are created in parallel with relation to each other and in the same time span with interlinking planned from the very beginning. A different case is the linking to previously published lexical resources, namely the LSP dictionaries *Social Work & Social Pedagogics*, *Health & Nursing*, and *Horticulture & Landscaping*.

When comparing these to the *DGS Corpus* and *DW-DGS*, several important differences can be observed:

- They cover specialised language and were aimed at sign expressions of technical terms as opposed to everyday language in *DGS Corpus* and *DW-DGS*.
- The main portion of the data collection involved elicitation of isolated signs for technical terms following a German word list as opposed to natural signing in context. Answers consist of a demonstration of the respective signs and do not include their actual use in a linguistic context, a prerequisite of analysing usage.
- Due to the elicitation method it was not always completely clear which of the answers were established signs and which were spontaneously made up translations such as loan translations, homophone calques and productive signs (cf. König et al., 2008, p. 380). For an evaluation and selection of the signs to be shown in the dictionaries, native speakers' intuition of Deaf team members and the recurrent use by several informants were used as criteria.
- Methodological and technical aspects of elicitation, annotation and production were according to the standards of the respective time. This means that the quality of contents and lemmatisation may be somewhat outdated in comparison to today's standards and rules.

Although the data of the LSP dictionaries are stored and maintained in iLex, it happened for several reasons that IDs used for type entries in the gloss index of an LSP dictionary changed or got lost. In these cases the IDs have to be reconstructed or a mapping with actual type IDs needs to be done manually.
For the joint German index the challenge was to come up with a feasible rule to filter out links to LSP sign entries that were already covered by *DW-DGS* entries.

4.2. Rationale for Linking to Older Resources

The *Public DGS Corpus* and *DW-DGS* are intended to become the preferred reference tools for information on DGS when finished. Since they are online products they can be interconnected with each other and with other lexical resources of DGS and can thus serve as a common gateway also to these other resources. Resources can be linked without too much extra cost when the technical matching of sign entries to the entries of the respective resources can easily be achieved, when there is no legal problem with access rights and it can be ensured that the other resources will be unchanged and stay available in the future (sustainability). All these conditions are fulfilled for the LSP dictionaries in question. Reasons for linking are:

- Linking from the *MY DGS – annotated* type entries to the LSP dictionaries can easily be achieved because of shared iLex type IDs.
- Sign entries of the LSP dictionaries contain descriptions and general information on the simplex signs that were used in translations for technical terms. These

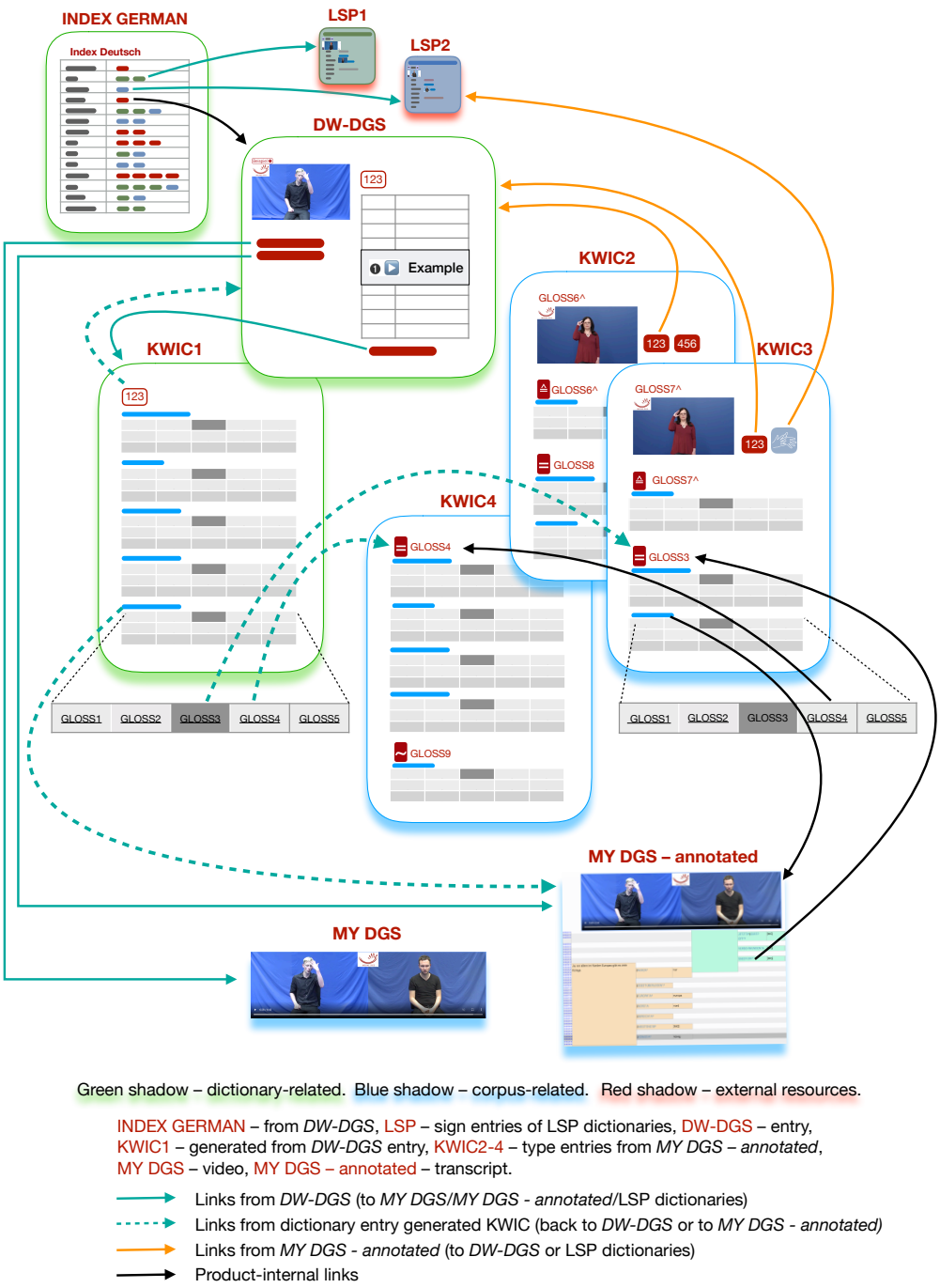

Figure 5: Implemented linking from corpus, dictionary, and other DGS resources.

signs were "[...] described almost as they would be in a general sign language dictionary" (König et al., 2008, p. 387). Entries include a representative movie of the citation form, identified conventional meanings and for iconic signs a description of the underlying image. This information serves the same information needs of the user as the *DW-DGS*, that is, information on the typical, everyday use of a specific sign.

- While the first entries of the pre-release *DW-DGS* are published online this resource should contain material on as many signs as possible so that a user can find at least some information when searching for a sign – even if there is not yet a fully finished corpus-based entry available. Including older information on signs that is already available and easily integrated into the resource increases the chances that a user finds useful information even at this early stage of production.

- LSP sign entries include a description of the iconic base of the signs, a piece of information not included in the *DW-DGS* entries. Making this information available can be considered as an additional gain. This is one of the reasons to link from the *MY DGS – annotated* type entries to LSP sign entries also in cases when a *DW-DGS* entry already exists.

There are two places where linking from DGS-Korpus products to the LSP dictionaries is implemented.

4.3. Linking from Corpus

MY DGS – annotated type entries link to LSP sign entries whenever a matching is available to one of the LSP products. The links are shown even if there is also a preferred link to an already existing *DW-DGS* entry. Links are done via a button representing the LSP dictionary and jump directly to the corresponding LSP sign entry (see box 'KWIC3' in Figure 5).

4.4. Linking from German Index of *DW-DGS*

The German index of the *DW-DGS* is compiled from translational equivalents provided in the entries for different senses of the described signs. German words with disambiguating context link directly to the corresponding sense in the respective entry. Not all equivalents given in the entries appear in the index. More systemic equivalents are included while less systemic equivalents (Hausmann and Werner, 1991; Héja, 2017) are excluded to avoid confusion. For those that are to appear in the index disambiguating information is added whenever the need arises to differentiate between separate senses of the German word or to distinguish between different sign senses to which the equivalents are addressed. LSP dictionary sign entries include one or several conventional meanings of the sign, realised as a German word translation sometimes with a disambiguating context added. These equivalents and contexts can be used to produce a joint German index of *DW-DGS* and LSP sign entries. *DW-DGS* translational equivalents and their disambiguating contexts are controlled for consistency while LSP translational equivalents and contexts come as they are in the product. In order to lead users to the preferred source of information – that is the corpus-based *DW-DGS* – and to avoid the confusion of multiple entries covering roughly the same scope only links to LSP sign entries are given when there is not yet a *DW-DGS* entry available.

When there is no disambiguation context given for the LSP equivalent but already existing, disambiguated *DW-DGS* equivalents, the links to the LSP sign entries are filtered out to avoid confusion and because the expectation is that *DW-DGS* sense covering might just be more detailed. However, this automatic filtering as a consequence might also filter out links to additional signs covering the same concepts or additional senses of the German word not contained in the *DGS Corpus* material and therefore not covered by the *DW-DGS* entry. In order to avoid taking out links to material not covered by the *DW-DGS* entries a manual inspection of possibly conflicting cases would be necessary to decide each case individually.

The resulting joint German index includes German words with or without a disambiguating context and links to either the *DW-DGS* entries or to sign entries of one or several LSP dictionaries (see box 'INDEX GERMAN' in Figure 5). Links to a *DW-DGS* entry appear as a red button with entry number and sense number, links to LSP entries are shown as IDs.

5. Conclusion

The DGS-Korpus project meets the vision of Kristoffersen and Troelsgård (2012, p. 99) of integrating sign language corpora and co-built dictionaries in some points. A combined product combines benefits of both a dictionary and a corpus, in addressing different user groups in various ways, providing independent use of either resource, but also close interconnection. Thus it respectively invites the language community or linguists to benefit from either the corpus or the dictionary.

With stand-alone products, there is no need to intermediate the scope of dictionary entries and the scope of type entries. In addition, as only the annotated corpus uses glosses, there is no conflict of labels. But the point of possible confusion has shifted to the places where dictionary and corpus are interlinked (see Section 3.2). This drawback is, in our view, clearly outweighed by the advantages: The interlinking documents how *DW-DGS* and *MY DGS – annotated* are built upon the same basis in a transparent way, it supports full access to resources and offers a large pool of usage examples.

Asmussen (2013, p. 1084) sets a high standard in the kind of interrelationship of what he calls a "combined dictionary-corpus product in the strict sense": Dictionary and annotated corpus "should be separately accessible" and "they should be linguistically interlinked, i. e. syntactically, semantically, and that means not only by shallow string similarities." He suggests a sense-specific linking of corpus tokens to dictionary entries (Asmussen, 2013, p. 1086). From what has been said above, a sense-tagging of the complete annotated sign language corpus is not feasible within a reasonable time. Instead, we offer a way to access from a corpus token via the referenced type to the dictionary entries. Users are able to scan the sense overview in the entry and check against the given sense definitions. For the future prospect, we think a crowd-sourcing tool that engages users to allocate tokens to the best fitting sense of the corresponding dictionary entry would be useful. These feedback inputs could be gathered, evaluated and redelivered in order to enhance the quality of KWIC concordances.

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 Academies of Sciences and Humanities.

7. Bibliographical References

Asmussen, J. (2013). Combined products: Dictionary and corpus. In *Dictionaries. An International Encyclopedia of Lexicography – Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, Handbooks of Linguistics and Communication Science, pages 1081–1090. De Gruyter Mouton, Berlin, Boston.

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., Konrad, R., and Schwarz, A. (2001). GlossLexer: A multimedia lexical database for sign language dictionary compilation. *Sign Language & Linguistics*, 4(1-2):171–189.

Hanke, T., Schulder, M., Konrad, R., and Jahn, E. (2020). Extending the Public DGS Corpus in Size and Depth. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, Marseille, France. 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.

Hausmann, F. J. and Werner, R. O. (1991). Spezifische Bauteile und Strukturen zweisprachiger Wörterbücher: eine Übersicht. In *Wörterbücher: Ein internationales Handbuch zur Lexikographie*, Handbücher zur Sprach- und Kommunikationswissenschaft, pages 2729–2769. De Gruyter Mouton, Berlin, Boston. Reprint 2017.

Héja, E. (2017). Revisiting Translational Equivalence: Contributions from Data-Driven Bilingual Lexicography. *International Journal of Lexicography*, 30(4):483–503.

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., and Langer, G. (2008). What's in a Sign? Theoretical Lessons from Practical Sign Language Lexicography. In *Signs of the Time. Selected Papers from TISLR 8*, pages 379–404, Barcelona, Spain. Signum-Verlag. The International Conference on Theoretical Issues in Sign Language Research took place at the University of Barcelona between 30 September and 2 October 2004.

Konrad, R. and Langer, G. (2012). Fachgebärdenlexikographie am Institut für Deutsche Gebärdensprache. *eDITion – Fachzeitschrift für Terminologie*, 1/2012:13–17.

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.

Konrad, R. (2011). Die Erstellung von Fachgebärdenlexika am Institut für Deutsche Gebärdensprache (IDGS) der Universität Hamburg (1993-2010). Revised version of doctoral thesis.

Kristoffersen, J. H. and Troelsgård, T. (2012). Integrating corpora and dictionaries: Problems and perspectives, with particular respect to the treatment of sign language. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 95–100, Istanbul, Turkey. European Language Resources Association.

Langer, G., Troelsgård, T., Kristoffersen, J., Konrad, R., Hanke, T., and König, S. (2016). Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages at LREC*, pages 143–152, Portorož, Slovenia. European Language Resources Association.

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.

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.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge University Press, Cambridge, United Kingdom.

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/`.

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/`.