

Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing

Gabriele Langer¹, Thomas Troelsgård², Jette Kristoffersen²,
Reiner Konrad¹, Thomas Hanke¹, Susanne König¹

¹Institute of German Sign Language (IDGS), University of Hamburg, ²Center for Sign Language, UCC
E-mail: {gabriele.langer, reiner.konrad, thomas.hanke, susanne.koenig}@sign-lang.uni-hamburg.de,
{ttro, jehk}@ucc.dk

Abstract

In a combined corpus-dictionary project, you would need one lexical database that could serve as a shared “backbone” for both corpus annotation and dictionary editing, but it is not that easy to define a database structure that applies satisfactorily to both these purposes. In this paper, we will exemplify the problem and present ideas on how to model structures in a lexical database that facilitate corpus annotation as well as dictionary editing. The paper is a joint work between the DGS Corpus Project and the DTS Dictionary Project. The two projects come from opposite sides of the spectrum (one adjusting a lexical database grown from dictionary making for corpus annotating, one building a lexical database in parallel with corpus annotation and editing a corpus-based dictionary), and we will consider requirements and feasible structures for a database that can serve both corpus and dictionary.

Keywords: corpus building, sign language corpora, lexicography, sign language lexicography, annotation tools, lemmatisation, sense discrimination, dictionary writing system

1. Introduction and Backgrounds

The German Sign Language (DGS) Corpus Project and the Danish Sign Language (DTS) Dictionary Project both have the aim to work on corpus-based lexicography within the iLex environment and want to have dictionary and corpus information in the same database.

Some signs are highly polysemous and have many phonological variants, while others have lots of variants and only a few senses, and yet others have only one form, but lots of senses. This is no problem in a database that serves exclusively as a type inventory for corpus annotation dealing with language documentation and description, yet it causes trouble for the lexicographer, who often needs to work with both flexible and in some cases pragmatic principles in order to produce dictionary entries that are human-readable as well as fairly homogeneous in appearance. In this paper, we will exemplify the problem and present ideas on how to model structures in a lexical database that facilitate corpus annotation as well as dictionary editing.

1.1 DTS Dictionary

The DTS Dictionary (Center for Tegnsprog 2008-2016; Kristoffersen & Troelsgård 2012) is a general-purpose dictionary describing the basic sign vocabulary of DTS. The dictionary has search facilities allowing for lookups based on sign form, Danish equivalent or topic (or a combination of these).

The core of the dictionary-making process is a semantic analysis of each selected sign – a task that so far has been performed partly based on introspection by staff members who are native signers, partly based on evidence found in video recordings. As the DTS group is now starting a corpus project, the aim is to build a tool that will on one hand facilitate the editing of new sign entries, and on the other hand supply tools for “retro-corpus-basing” existing entries, e.g. by checking

for missing word-senses, retrieving collocation information, or finding better usage examples. Finally, a corpus will be an essential tool in connection with future lemma selection, and could be used for other linguistic research outside the lexicographic context.

The aim is, as an obvious starting point, to re-use the sign lemmas, which are already uniquely glossed, as the core type vocabulary for the token-type matching during annotation of corpus texts, adding new signs along the road. Furthermore, the aim is to also exploit the word-senses defined in the dictionary entries, so that a token – if a suitable sense is at hand – can be matched directly to a sign type with a specific meaning.

1.2 DGS Corpus

The DGS Corpus Project is a long-term project with two major aims: building a reference corpus for DGS¹ as a multi-purpose resource for research on DGS and compiling a general dictionary of DGS on basis of the corpus data collected. Coming from a background of compiling German – DGS language for specific purposes dictionaries (1993-2010)² the annotation tool and integrated lexical database iLex has been developed in the context of these previous projects to facilitate the lemmatisation and annotation of recorded signed data (Hanke 2002; Hanke & Storz 2008). This database containing type entries and lemmatised sign data from previous projects has been carried over and is being used and further developed alongside with the iLex program to suit the needs of the DGS Corpus Project.

In the first stages of the project the focus has been on data collection and annotation, the latter will continue for several years to come to provide the data for general re-

¹ A representative part of the data is published from 2015 on as a subcorpus (DGS Corpus Project, 2015-2016).

² For more information on the LSP dictionaries cf. Konrad (2011) and Konrad & Langer (2009).

search and lexicographic analysis. Thus, though the Corpus Project aims at compiling a dictionary as well as building a corpus, up to now the main focus on the development and use of iLex has been on annotation rather than lexicographic description. In the near future iLex structures have to be developed further to better support analysis, the various stages of working out lexicographic descriptions of sign uses as well as the writing of dictionary entries.

2. iLex

The iLex program is a database and annotation environment developed at the Institute of German Sign Language (IDGS, University of Hamburg,) especially for annotating sign language data. Within the annotation environment of iLex video files can be viewed and tagged as in other annotation tools like e.g. ELAN (cf. Crasborn & Sloetjes 2008), where media and annotations are time-aligned. Unlike ELAN, iLex combines a lexical database with transcript views for annotation of video segments. Lemmatisation as a process of identifying tokens as instantiations of sign types (token-type matching) is done by establishing a direct and dynamic link between type and token via drag & drop. Thus, consistency is supported by the database structure and does not rely on the use of ID-glosses (see Johnston 2010).³ Type entries include information on the presumed lexical types and allow direct access to all tokens of the respective type. Furthermore, iLex provides a type hierarchy with several levels that allows modelling relevant differences in iconicity, form, and use of a sign and tagging the tokens accordingly. Each type of a lower level is attached to exactly one type of a higher level and is considered to represent a subset of the tokens and uses belonging to the superordinate type.

2.1 Use of iLex Structures by DGS Group

In the DGS Corpus Project, four type levels are being utilised. A type at level 3 – hereafter called *supertype* – represents a sign as an abstract linguistic entity (with focus on form and – if existent – iconicity). Types at level 1 – hereafter called *subtypes* – are defined to distinguish different established or conventional uses of a sign with regard to meaning. Conventional uses of a sign typically consist of regular and therefore expectable sign-mouthing combinations. A subtype is directly attached to its supertype if they share the same citation form.

Tokens that show productive or novel uses of a sign or not yet identified conventional uses are matched to the supertype directly. Productive uses are for example occasional or ad-hoc sign-mouthing combinations (cf. König et al. 2008, 398-400).

With the implementation of qualifiers (Konrad et al. 2012), also word forms and other form differences within one supertype or subtype can be classified and

labelled. For this purpose the type levels 2 (qualified supertypes) and 0 (qualified subtypes) were introduced. Qualified types allow distinguishing and coding modified forms and also “minor variants” (Johnston 2016, 19-20) of the sign form as part of a more detailed analysis of a sign’s use.⁴ The goal of this coding is to determine the range of form variation and modification within the given supertype or subtype.⁵ On the subtype level, qualified forms (level 0) are either candidates for word forms or phonological variants or they may be just performance phenomena. On the supertype level form variation (modelled by level 2) can be cases of modification, phonological variation, derivation or performance phenomena.

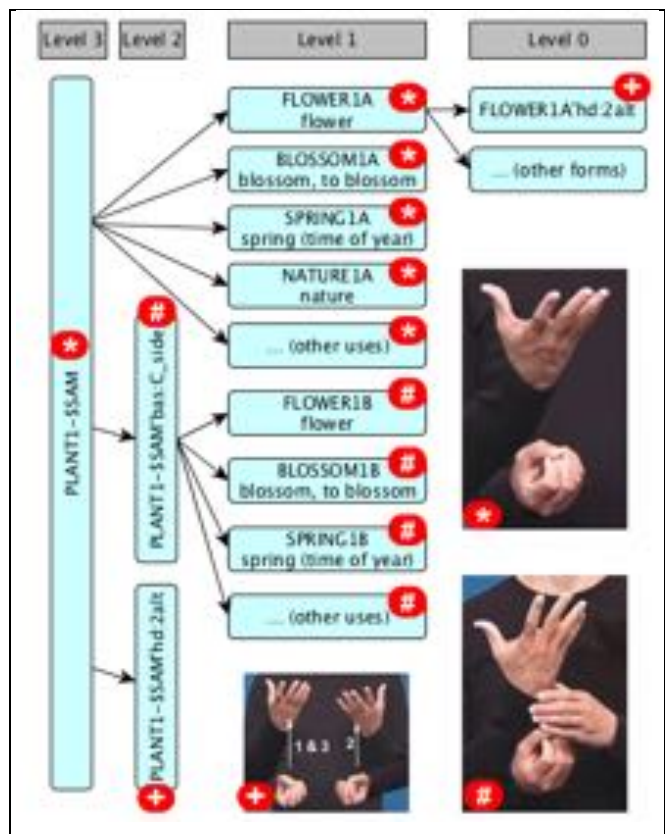


Figure 1: Type hierarchy of DGS sign PLANT1-\$\$AM⁶

Supertype entries are considered to be lexical entities whereas subtypes group together conventional sign uses, often triggered by mouthing. Roughly speaking, the supertype and subtype structure is used to model polysemy. For lexicographic descriptions the conventional sign uses have to be further analysed for different senses

³ The iLex database uses type IDs for identification and linking. However, for the ease of human annotators each type is assigned a unique gloss in the database that functions like an ID-gloss and as a mnemonic aid e.g. when reading transcripts or referring to signs in communication. iLex blocks attempts to name a new type entry with a gloss already used and thus makes sure that glosses are unique on each level (cf. 2.1).

⁴ This coding is not part of the basic annotation. It can be done completely for all tokens of a sign or selectively at a later annotation pass (lemma revision or detailed transcription).

⁵ The focus here is on the individual types because of the lexicographic perspective. However, coding the same modifications or form deviations across different types in the same way will also allow to run analyses across a number of types.

⁶ In order to distinguish supertypes from subtypes, glosses of supertypes always have the suffix “-\$\$AM” (abbreviation of ‘Sammelglosse’ (collective gloss)).

of a sign (sense discrimination).

Figure 1 exemplifies the type hierarchy structure in iLex as it is used by the DGS group. In the lexical database everything that presumably belongs to one sign is in some way hierarchically attached to the same supertype. Depending on their form and contextual meaning, tokens can be attached to types on each level. Tokens attached to types on lower levels are always considered to be at the same time instantiations of the superordinate types (“double glossing”). Supertypes and subtypes have regular glosses while qualified types have codes and values attached to the gloss of the superordinate type.

2.2 Use of iLex Structures by DTS Group

The DTS group already has dictionary entries with a semantic differentiation and intends to re-use these entries as sign types for annotation. For this purpose they have imported part of their dictionary entry structure into iLex.

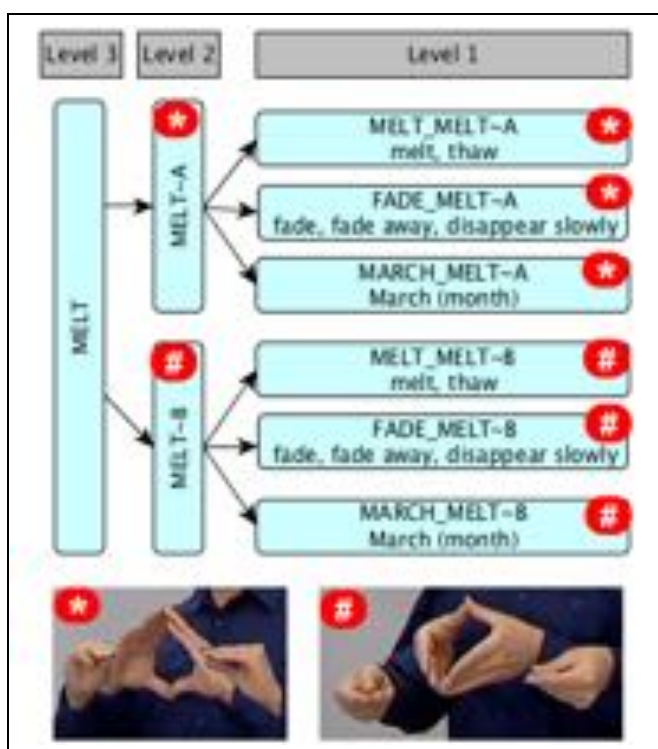


Figure 2: Type hierarchy of the DTS sign MELT

In the DTS Dictionary one sign entry may include different variant forms (one of which functions as citation form of the sign) and several senses of a polysemous sign. Three type levels have been used to model this structure in iLex. Types at level 3 (supertype) represent the whole lemma sign and therefore the whole entry. The different forms of a sign (GLOSS-A = citation form, GLOSS-B = variant) are represented as types at level 2. These form types are attached to the supertype. Subtypes (level 1) are used to represent the different senses as conventional uses of the sign in its respective variant form (see figure 2).

2.3 Differences in the Use of iLex Structures

The DGS and DTS groups use iLex structures in a quite similar way: Supertypes (level 3) represent the lemma sign, types of level 2 distinguish form variations, and subtypes (level 1) represent different conventional uses of the sign form (i.e. roughly meanings (DGS) or senses (DTS)). However, the DGS group does not repeat the supertype form (technically functioning also as citation form) on level 2 but links conventional uses of that form directly to the supertype, whereas the DTS group has all form types on level 2 (replicating the citation form) and therefore no direct linking from subtype to supertype. Apart from glossing conventions for phonological variants, the DGS group codes form types via qualifiers (and their values) to categorize form differences with regard to the citation form across types, while the DTS group does not use qualifiers at the moment.

The iLex database does not dictate how the type structures should be used, and in addition to the two models described above, the system can be designed to work with any model from a one-dimensional type list to a complex multi-level structure.

3. Lexicographic Needs

3.1 Corpus Data as Basis for Sign Description

Annotated sign language (SL) texts should serve as the basis for the different kinds of analyses performed during the lexicographic description of a sign, e.g. establishing overviews of phonological variants and modified forms, of meanings, of usage (collocations, grammatical functions), or of distribution (with regard to region, age, gender etc.). Therefore, tokens that are likely to end up in one dictionary entry (or in one sense) should be tagged uniformly during corpus annotation.

Furthermore, the corpus system should provide tools for performing these analyses, e.g. tokens in context-view [concordance view], frequency lists, collocation statistics (Mutual Information, T-score etc.). In addition, the system should facilitate access to information from outside the corpus itself, e.g. data from informant surveys.⁷

3.2 The Lexicographic Workbench

When determining the final meanings' structure of a dictionary entry, it is good practice first to get an overview by describing the occurring senses at a rather high level of detail, and only in a second step to lump together closely related senses, preferably preserving the preliminary, fine-grained analysis to be consulted in connection with later revisions of the entry (cf. Atkins & Rundell 98-101, 268). The ideal integrated corpus-dictionary system should hence accommodate a preliminary, “full” set of senses, as well as a “cleaned-up” set that constitutes the meanings' structure of the final entry. Furthermore, these sets should be linked together

⁷ The DGS Corpus Project uses an online survey called *DGS Feedback* to gather further information on signs and their use (cf. Langer et al. 2014, Langer et al. 2016). The results of this survey are complementary to corpus data and should be easily accessible when making lexicographic analyses.

in order to keep track of to where which senses go, and to more easily change the structure at later revisions. On both processing stages, there should also be place for storing further information regarding the decisions made during the analyses, e.g. hypotheses, questions, and comments.

In addition to this, a joint corpus-dictionary database should obviously accommodate all information kinds one would like to have in the dictionary's entry structure, as well as the needed meta-data such as markers for status, workflow control etc.

3.3 Corpus Data as Empirical Evidence

A dictionary entry is a sort of claim of giving an accurate description of the use of a sign, and links to actual corpus occurrences will provide accessible evidence for these claims. Thus, the system should allow for linking from each sense or grammatical or pragmatic function described in the dictionary data to corpus data, both on the higher levels (types in the lexical database for annotation), and on the lower (specific token or phrase tags in the annotations). Similarly, base form, variants and modified forms shown in the dictionary can be supported by evidence via links to corpus occurrences or links to types on various levels.

For some dictionaries, you might want to present authentic usage examples to the users. These could be taken directly from the corpus videos, or they could be adapted and re-recorded, e.g. for anonymisation reasons, or for making the examples more accessible for L2 learners (cf. Kristoffersen, 2010). In both cases, there will be a need for linking from a particular place in the dictionary data to a corpus occurrence. If example sentences are re-recorded, you might incorporate these recordings into the corpus system as a separate sub-corpus, in order to be able to link from the relevant dictionary sense both to the original source, and to the final version of the sentence.

4. Corpus Needs

4.1 Annotation and Lemmatisation

In addition to translation, the core task of basic annotation of SL texts is lemmatisation (cf. Johnston 2016, 13-48), also called token-type matching. Here the focus – the first criterion for matching – is on form, meaning being secondary and only rather roughly distinguished. As lemmatisation is very time-consuming it is essential that the annotator can find and identify relevant types as easily and fast as possible and with a reliable result.⁸ One prerequisite for this is access to the up-to-date state of lexical entries (type entries). The system should also provide a number of easy-to-do searches via form, gloss, meaning, mouthings and combinations thereof across type entries and already lemmatised tokens.

⁸ The DGS Corpus database contains several thousands of type entries. In order to be able to find and identify the right supertype or subtype effective search strategies are necessary. iLex supports the lemmatisation by searches for and easy entering of the correct types into the transcript in various ways. When the annotator finds a good supertype candidate for token-type matching, the type hierarchy allows for getting a quick overview of the range of form and meaning aspects connected with one type to choose the best match.

For a fast check whether the found type is the correct one, the system should provide easy access to the citation form of the sign – for example by offering a representative video clip to be played (either a studio recording or an already lemmatised representative token), and also provide fast access to other tokens of that type for comparison. Also, when there is no fitting type to be found, annotators should be able to add a preliminary new type to the system.

Annotators should not be left in doubt what to do with tokens that are unusual with regard to their contextual meanings (productive uses or not yet identified conventional uses) or that are ambiguous in their meaning. The annotation conventions should cover these cases, and ideally the annotation tool should provide a mechanism to link them to a suitable type and at the same time keep them separate for further analysis, as it is the case when attaching all these tokens directly to the supertype. In this way annotators do not need to brood over meaning differences and the discrimination of various senses in the process of basic annotation.

Depending on lemmatisation rules it may be the case that two or more supertypes entries share the same citation form (homophony). In these cases, if it is unclear to the annotator which of these supertypes to choose, any of the possible supertypes could be regarded as suitable in the first annotation pass, and the decision of choosing a more specific type could be deferred to a later stage⁹, see 4.2 below.

4.2 Lemma Revision

In order to insure consistency and quality of the lemmatisation, the DGS group found it helpful to establish a step they call *lemma revision* (cf. Konrad & Langer 2009). Here the focus is shifted from sequential text annotation to the single supertype and its forms and to some degree meanings. The token-type matching is checked in comparison to other tokens and the citation form. The tokens attached to the supertypes (productive uses) are checked for repeated occurrences of use in order to identify further conventional uses and establish new subtypes. The type structure is reviewed in the context of other types with related and similar forms and also taking into account sets of variants and modification behaviour. If necessary, the type structures are corrected or expanded. At the same time sign forms (modifications and variants) can be further distinguished (detailed annotation: levels 0 and 2). Cross-references between similar types are added. The result of the lemma revision is then a good basis for the ongoing lemmatisation. An annotation tool should allow one to conveniently access and collectively view all tokens of one (supertype) sign and compare them looking from different perspectives (form variation, meaning, relations to other signs etc).

5. Divergent Structural Needs

The type structure is the result of the lemmatisation process (including the lemma revision, cf. 4.2). Ideally, it

⁹ As the type hierarchy in the iLex database is flexible, another approach to solve the supertype homophony problem could be to introduce a formal, "form only" type above the supertype level, to be chosen when in doubt.

facilitates not only the ongoing annotation of corpus texts but also the building of a pre-structure of the data to be used as a basis for linguistic research, including lexicographic analysis and description of signs, their forms and uses. Lemmatisation in annotation focuses mainly (but not exclusively) on form in order to collect all instances of one sign under one label (be it an ID-gloss or an ID in a database), while establishing lemmas in lexicography focuses much more on the meanings of signs and has to consider additional factors (see 5.4). A rather complex example is illustrated in Fenlon et al. (2015, 196-198) where seven form variants are grouped into four separate lexemes in the BSL Sign Bank.

As implied above, the different sets of requirements might not always be fully compatible. Incompatibility problems may be due to annotation guidelines focussed on facilitating type search, to a clearer picture of the situation (e.g. different modification behaviour or different sets of variants) after analysis, or to pragmatic lexicographic decisions, e.g. keeping entries from becoming too large, complicated or counter-intuitive for the user. As a result, one will most likely in a number of cases end up with a different division or grouping of the data and types into separate or collective dictionary entries, thus diverging from the pre-structure built during the annotation.

In the following we will have a closer look at some of the factors that lead to structural incompatibility between corpus and dictionary, and at some concrete examples. Finally, we will shortly look at one of the challenges arising when combining both structures in interrelated products.

5.1 Ongoing Changes of Type Structures

Annotation is to be seen as an ongoing process of multiple passes each adding to and correcting the annotation while lemmatising or analysing or using the annotated data. In principle, this work is never finished as each look at the data with different goals or research questions in mind reveals new insights and can lead to an enrichment of the annotation.

In the DGS Corpus Project, up to 40 people are working at the same time in iLex on the data. New types are added as needed and type structures may be changed, expanded, rearranged or split as it seems necessary in the light of new data or analyses. Each change is available in real-time to all others. Therefore, type structures are at least in principle constantly changing.

At one point in time the scope of a dictionary entry has to be defined and fixed in order to be able to summarise and describe the data available. We are convinced that for this purpose we need a structure that allows us to do that and to fill in our information on this subset of data in a structured way without imposing dictionary writing decisions back onto the annotation database.¹⁰

¹⁰ However, sometimes dictionary analyses and decisions may reveal problems in the annotational type structure of a sign and may lead to a restructuring of it as well. Our point here is not that this should not happen – in fact this happens a lot in the DGS Corpus Project and is even an appreciated way of quality control – but that the system should be flexible enough to have the freedom of diverging groupings of data for annotation and

5.2 Two Supertypes – One Dictionary Entry

Consider the two type structures of WORK2-\$\$SAM (1) and END1-\$\$SAM (10) in the annotation database (fig. 3).

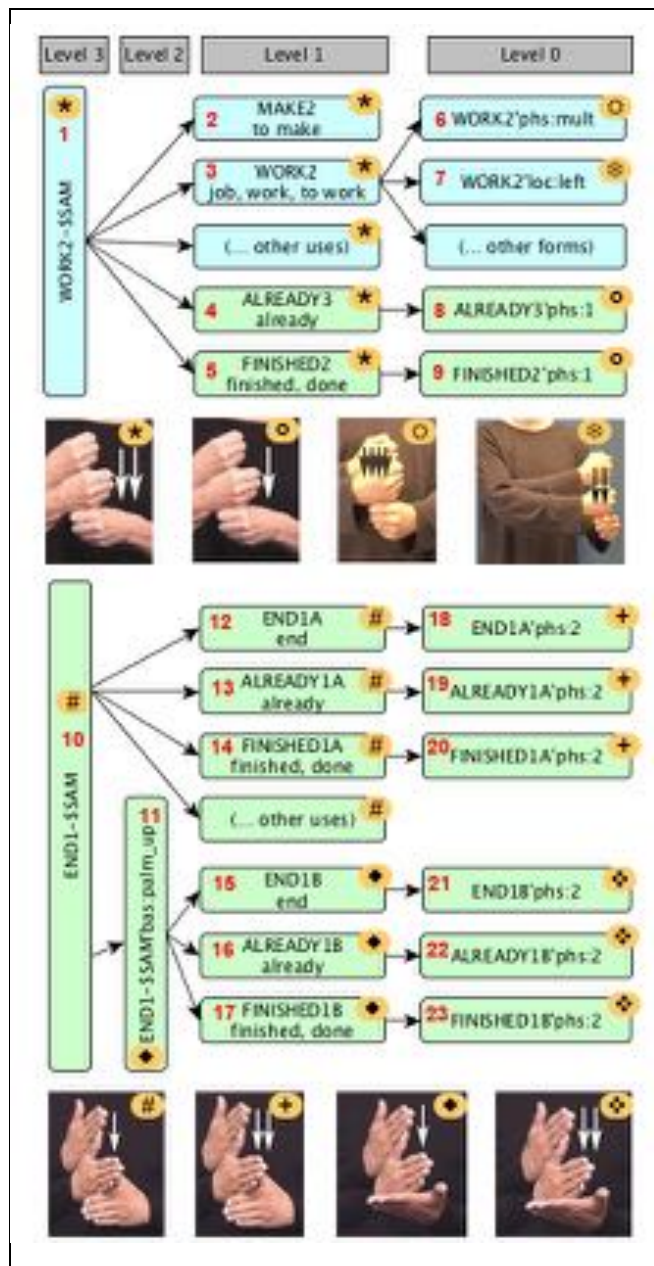


Figure 3: Type hierarchies of WORK2-\$\$SAM and END1-\$\$SAM (DGS)

The green parts of the WORK2-\$\$SAM type structure are completely fitting with regard to form and one can also easily see a semantic closeness of the different uses of WORK2-\$\$SAM which include among others ‘to work’ (3), ‘to make’ (2), ‘done/finished’ (5), ‘already’ (4) along the lines of ‘work that has been done is already finished’. For annotation this structure based on form as a first criterion is very convenient. However, the blue subtypes (‘work’, ...) and the green subtypes (‘finished’, ...) differ with regard to variants and modification behaviour, a fact that can be noticed when reviewing all the available data after lemmatisation. Another entry

lexicography.

WORK1-\$\$AM (circular movement to hit the base hand repeatedly) is very similar in form to WORK2-\$\$AM and also shares most of its meanings in the *work* domain, but not the senses ‘already’ and ‘finished, done’. Both signs share modification behaviour, for example they can be used in an intensified form with a smaller movement including several repetitions (6). On the other side there is END1-\$\$AM with its two variant forms (10) and (11) being very similar to the green parts of the WORK2-\$\$AM structure. (10) and (11) differ only with respect to the parameters handshape and \pm repetition. These forms could be considered phonological variants on basis of similarity of form and meaning, especially since they both can be made with repetition (4, 5, 18-20) and without repetition (8, 9, 12-14). This is also true for the third form (11), which differs from (10) only in the orientation of the base hand. Considering the differences in modification behaviour of WORK2 (3) and other uses in the *work* domain and ALREADY3 (4) and FINISHED2 (5) as well as the formational similarity and semantic overlap of ALREADY3/FINISHED2 with the END1-\$\$AM sign uses, it seems reasonable and adequate to virtually re-group the material into two different dictionary entries WORK and FINISHED, as can be seen in figure 4.

WORK	FINISHED
<u>Citation form:</u> * \Rightarrow 1	<u>Citation form:</u> # \Rightarrow 10
<u>Variant:</u> [O ɛ ɹ ɔ ɜ ɔ] 11 C X +	<u>Variants:</u> ♦ \Rightarrow 11, ○ \Rightarrow 8, 9
\Rightarrow WORK1-\$\$AM	<u>Note on form:</u> All 3 forms can be made with repetition \Rightarrow 4, 5, 18-23
<u>Modification behaviour:</u> can be modified for loci \Rightarrow 7,
intensive: ⚙ \Rightarrow 6	Senses:
...	1. end \Rightarrow 12, 13
<u>Senses:</u>	2. finished, done \Rightarrow 5, 14, 17
1. work (task), to work \Rightarrow 3	3. already \Rightarrow 4, 13, 16
2. work, job \Rightarrow 3	4. ...
3. to make \Rightarrow 2	
4. ...	

Figure 4: Pre-dictionary entries¹¹ of WORK, FINISHED

5.3 One Supertype – Two Dictionary Entries

In the DGS group there is a tendency in the first pass of basic annotation to formationally group derived forms under the more common or basic sign form (supertype), even in cases where the lexicographer would arrive at the interpretation of two related but independent signs. An example for this is STAMP1-\$\$AM.

The sign type hierarchy of the sign STAMP1-\$\$AM (30, (see figure 5) would be split into two entries, one covering the blue and the other covering the green part of the structure. The common citation form (30) could be interpreted as an iconic resemblance of pressing a stamp

onto a sheet of paper (fist onto flat hand), and it has several conventional uses including: ‘civil servant’, ‘agency’, ‘patent’, ‘visa’, ‘authentication’, ‘to stamp’. It has a form variant (31) where the non-dominant hand is not a flat hand but also a fist.

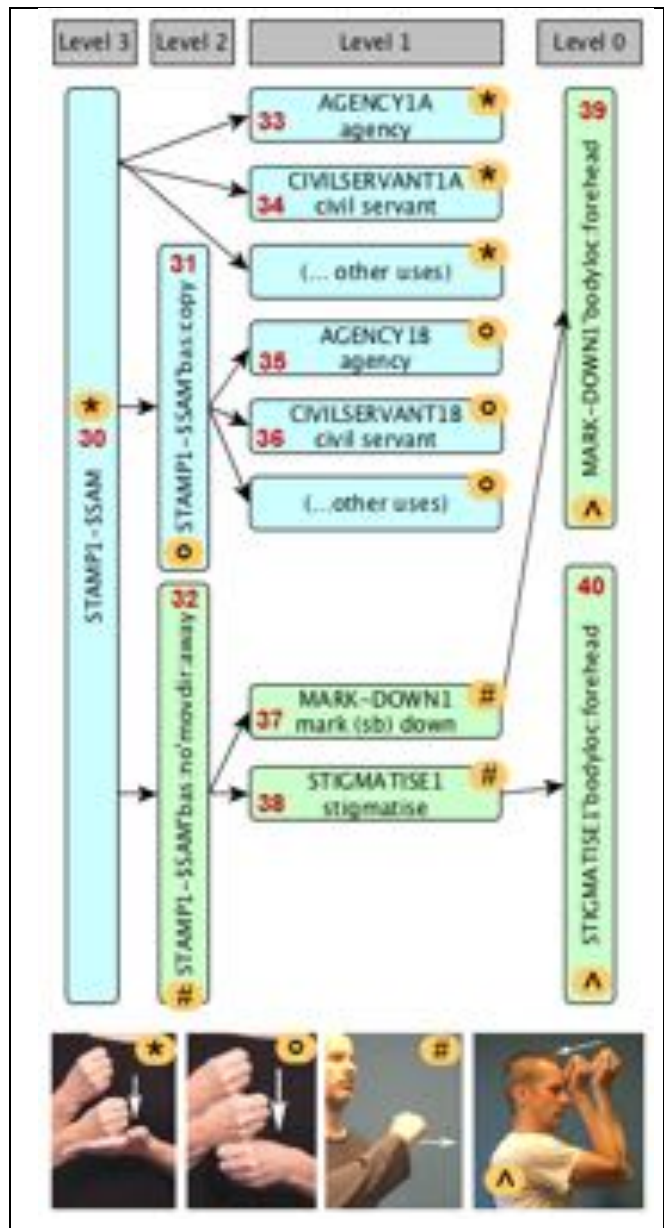


Figure 5: Type hierarchy of STAMP1-\$\$AM (DGS)

There is a derived form of the sign STAMP1-\$\$AM made with only one hand that moves forward instead of downward as if to stamp somebody else on the forehead (32). This form is conventionally used with the meanings ‘mark somebody down’ (37) or ‘stigmatize’ (38), both figurative uses of ‘stamp’. Both (37) and (38) presumably¹² can appear in context also in the form of stamping the signer’s own forehead (39, 40). Even though (32) is most likely derived from (30) it would get its own dictionary entry MARK-DOWN independent of the dictionary entry STAMP (see figure 6) because the uses of the sign differ with regard to form and meaning¹³ from

¹¹ Pre-dictionary entries are content of the pre-dictionary database as described by Atkins & Rundell (2008, 98-100) and therefore of the Dictionary Writing System module planned for iLex. Everything red in the entries refers to evidence in the corpus (types or particular tokens of that type in the annotation database) and could be instantiated by links to the corpus data and annotation types. Red numbers and orange codes for form refer to numbers and codes in figure 3.

¹² No corpus data yet to back up this claim.

¹³ Cf. Battison (2005, 240): “If two signs are made differently,

the other uses of the sign STAMP1-\$\$AM (e.g. 33-36) and also because it exhibits different modification behaviour.¹⁴

STAMP Citation form: ★ ⇒30 Variant: ○ ⇒31 ... Senses: 1. agency ⇒33, 35 2. civil servant ⇒34, 36 3. ...	MARK-DOWN Citation form: # ⇒32 Word forms: 1st person: ▲ ⇒39, 40 ... Senses: 1. mark somebody down ⇒37 2. stigmatise ⇒38
---	--

Figure 6: Pre-Dictionary entries¹⁵ of STAMP and MARK-DOWN (DGS)

5.4 Editorial Principles

The task of defining exactly what constitutes a dictionary lemma differs from the corresponding task performed in connection with annotation (lemmatisation). In lexicography, the focus is more towards the meanings of signs and has to consider additional factors, including even practical matters such as entry size and user-friendliness. Typically, dictionaries have their own “lemmatisation rules” describing which kinds of words (or signs) to include into the dictionary, which to give their own independent entries and which to treat as run-ons or sublemmas, specifying how the particular dictionary treats cases of homonymy and polysemy. Svensén calls this step the *establishment of lemmas* and distinguishes it from lemmatisation (cf. Svensén 2009, 94).¹⁶

The DTS group works with principles for establishing lemmas that are partially based on phonological, partially on semantic criteria (cf. Kristoffersen & Troelsgård, 2010). One criterion is that a figurative use of a sign will be described as a sense in the main entry only if the semantic relation is synchronically transparent. Otherwise, the figurative use will be established as a separate (homophone) sign entry. An example of this is a sign that can mean ‘red’ as well as ‘social’. As a consequence of the editorial principles of the dictionary, the sign is formally split into two entries, RED and SOCIAL (fig. 7).


RED Senses: red		SOCIAL Senses: social	
------------------------------	--	------------------------------------	---

Figure 7: Dictionary entries of RED and SOCIAL (DTS)

Another principle is that a sign described as having several variant forms can include only senses that can all be expressed through one particular variant (shown as the citation form in the dictionary). Any variant-specific sense gets its own independent sign entry in the dictionary. An example is the sign PRAY (see figure 8), with two variants, of which only one can mean ‘sorry’. Hence, although ‘sorry’ could be considered as semantically transparently related to ‘pray’ or ‘beg’ (= ‘beg for forgiveness’), it is established as a separate lemma.




SORRY Sign form:  Senses: sorry, pardon	PRAY Sign form 1: Sign form 2:   Senses: pray, prayer
--	---

Figure 8: Dictionary entries SORRY and PRAY (DTS)

These examples from the DTS Dictionary show that the needs of lexicographers and annotators are different due to their differences of viewing the same phenomena from different perspectives. It would be unwise to force the dictionary decisions onto the annotation database structure for reasons of resources but also because different dictionaries made of the same corpus lexical database / corpus may work with different lemma definitions targeting their specific user groups.

5.5 The Glossing Challenge

In spite of the incompatibility problems mentioned above, a corpus project and a dictionary based thereon could easily coexist if they were regarded as two independent language resources. However, having two such interlinked resources gives you some advantages that you would obviously like to pass on to the users, e.g. by showing corpus data directly in the dictionary, by linking from the dictionary to corpus occurrences of a sign or a sense, or by linking from corpus searches to relevant dictionary entries. And here – when presenting the resources as two interrelated products (or perhaps even as one integrated website) – the challenge arises of how to represent the signs in a user-friendly way. For a DGS sign entry like FINISHED (see figure 4), the dictionary entry as a whole cannot unambiguously refer back to the right corpus type gloss. The corpus on the other hand will contain many sign types that are not covered at all by the dictionary. If both products work with glosses as labels, using different gloss systems might be confusing for the user and a merging of the glosses might prove difficult.

and have different meanings, this is good evidence that they are separate signs.”

¹⁴ Actually, during lemma revision the branch starting with (32) was taken out of the type hierarchy of STAMP1-\$\$AM and got its own supertype entry also in the annotation database.

¹⁵ Red numbers and orange form codes refer to figure 5. See footnote 13 for further explanations.

¹⁶ “...ESTABLISHMENT OF LEMMAS. This operation is not simply the same thing as lemmatization... Naturally, the establishment of lemmas assumes previous lemmatization, but it also includes deciding how lexical items having identical base forms are to be presented in the dictionary, and to what extent word elements and multi-word lexical items are to be accorded lemma status.” (Svensén 2009, 94)

6. Conclusions

6.1 Structural Requirements

As we have shown above, a lexical database structure built for corpus annotation and one built for dictionary-making serve different purposes, and may therefore arrive at different suggestions for a structured description of the form and meaning of a particular sign. Where the lexicographic side typically is governed (to a higher degree) by semantic criteria, as well as by editorial rules and pragmatic decisions, the annotation side is mainly form-oriented, and should ideally facilitate that the lookup of a particular sign form always leads the user to the same type entry. Trying to perform a concurrent adjustment of the annotation base structure in order to match the dictionary structure at any time would eventually obscure the originally form-based structure, and hence hamper the lemma identification.¹⁷ We therefore suggest a model with different structures for annotation and dictionary editing, that each serves its primary purpose optimally on one hand, and on the other hand are sufficiently similar to be related to each other – and to be interlinked.

6.2 Linking Corpus and Dictionary

As shown in section 5, divergences between annotation principles and editorial principles can lead to a need of linking from one entry in the annotation database to two or more entries in the dictionary, as well as the other way round. Hence the suggested structure should allow for a rather free linking structure. Furthermore, linking will be needed between a series of places in the dictionary structure and different places in the lexical database used for annotation.

If we walk through the entry structure of a comprehensive SL dictionary, links to corpus evidence could be relevant in many places. On the entry level, you would probably like to refer to the sign itself, as evidence of its phonological variants, just as you typically would list the citation form and variants of the headword in a written/spoken language dictionary. These references could ideally be established as links to one or more types (on any level) in the type hierarchy of the annotation base, suggesting that the sign in question matches the types perfectly. Knowing that in many cases this will not be possible, we suggest that the structure allows for linking not only to types, but also to individual tokens in the transcripts.

Similarly, where a written/spoken language dictionary shows inflected forms of the headword, you might want to show frequent modifications of the sign, with no

regard to meaning (linking to level 2 types) or to a group of tokens being instantiations of this modification (linking to level 0 types).

On the meaning level, links to evidence for the described senses would obviously be desirable and appropriate, as they document the performed semantic analysis. This documentation could be valuable not only from the lexicographic perspective, e.g. in connection with later revisions of the dictionary (based on an updated and extended corpus), but also from other perspectives, e.g. in connection with other linguistic research based on the corpus. Just as the linking on entry level, this linking could have either a type (all tokens of that type are covered by that sense) or a single token as target in the annotation database.

On the meaning level, you might also need evidence of particular modified forms of the signs, used in a particular sense. Again, linking both to type and token could be needed.

On the meaning level, it would also be obvious to place links to good usage examples found in the corpus (as well as to reproductions of these, cf. section 3.3). Evidence of frequent collocations with the sign would also be a linking candidate. These latter types of linking could be done to a phrase/utterance in a corpus transcript, instead of to a single token. Figure 9 shows a linking model that accommodates the basic needs as described above.

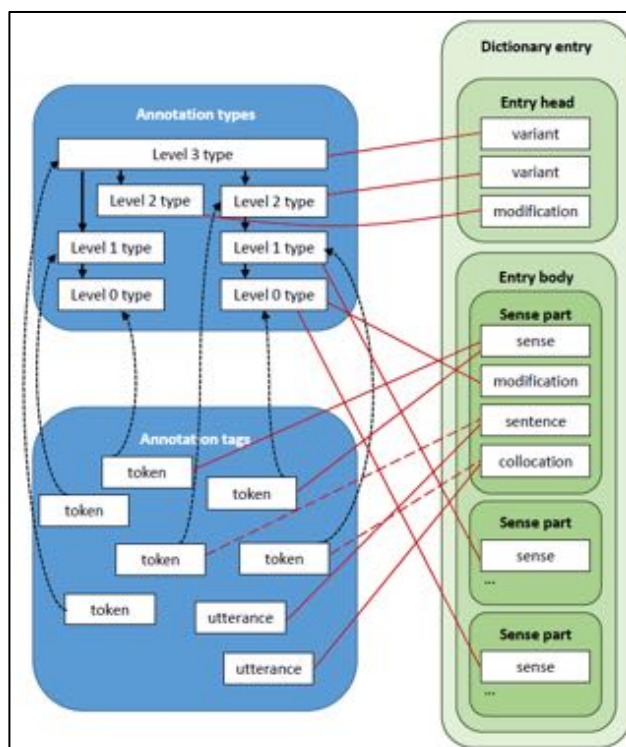


Figure 9: Suggested model for linking between corpus and dictionary

6.3 Future work

In the future, the two project groups will continue the work, aiming at adapting the iLex system to accommodate dictionary data, and to facilitate linking

¹⁷ In the LSP dictionary projects in Hamburg, up to 2010, sign entries had been produced directly from the iLex type entries via an export routine. Prior and close to the production, changes to iLex type entries had to be halted and in some cases type structures had to be adapted to suit the intended outcome as dictionary entries. Based on this experience the DGS group thinks it advisable to have a separate structure for the preparation of dictionary entries. Atkins & Rundell (2008, 98-100) also make a strong case for such a structure, which they call *pre-dictionary database*. Ideally, this structure should be part of the iLex environment and be interlinked with the annotation types.

between these data and appropriate types and tokens in the corpus data, as outlined above.

7. Acknowledgements

(Hamburg Group) 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.

(Copenhagen Group) The work at the DTS Dictionary project is supported by funding from the "Diversity and Social Innovation" research fund of UCC (Professionshøjskolen University College Capital, Denmark), from Alfred Jacobsens Foundation (Denmark), and from the Jascha Foundation (Denmark).

8. Bibliographical References

- Atkins, B.T.S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Battison, R. (2005). Signs Have Parts. A Simple Idea. In C. Valli, C. Lucas, & K.J. Mulrooney (Eds.), *Linguistics of American Sign Language. An Introduction*. 4th ed. Washington, DC: Gallaudet University Press, pp. 230--241.
- Crasborn, O., Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood, & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*. ELDA, Paris, pp. 39--43.
- Fenlon, J., Cormier, K., Schembri, A. (2015). Building BSL SignBank: The Lemma Dilemma Revisited. *International Journal of Lexicography*, 28(2), pp. 169--206. doi:10.1093/ijl/ecv008.
- Hanke, T. (2002). iLex. A tool for Sign Language Lexicography and Corpus Analysis. In M. González Rodríguez, & C. Paz Suarez Araujo (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. [Online resource; URL: <http://www.lrec-conf.org/proceedings/lrec2002/pdf/330.pdf>; last access: March 22, 2016]
- Hanke, T., Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood, & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*. Paris: ELRA, pp. 64--67.
- Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1), pp. 106--131.
- Johnston, T. (2016). Auslan Corpus Annotation Guidelines. February 2016 version. [Online resource; URL: http://www.academia.edu/12360442/Auslan_Corpus_Annotation_Guidelines_February_2016_version; last access: March 22, 2016].
- König, S., Konrad, R., Langer, G. (2008). What's in a Sign? Theoretical Lessons from Practical Sign Language Lexicography. In J. Quer (Ed.), *Signs of the Time: Selected Papers from TISLR 2004*. Hamburg: Signum, pp. 379--404.
- Konrad, R. (2011). *Die Erstellung von Fachgebärdenslexika am Institut für Deutsche Gebärdensprache (IDGS) der Universität Hamburg (1993-2010)*. Universität Hamburg. [Online resource; URL: http://www.sign-lang.uni-hamburg.de/projekte/mfl/konrad_2011_fachgeblexika.pdf; last access: March 22, 2016].
- Konrad, R., Hanke, T., König, S., Langer, G., Matthes, S., Nishio, R., Regen, A. (2012). From form to function. A database approach to handle lexicon building and spotting token forms in sign languages. In O. Crasborn, E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Interaction 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. [Online resource; URL: http://www.lrec-conf.org/proceedings/lrec2012/workshops/24.Proceedings_SignLanguage.pdf; last access: March 22, 2016].
- Konrad, R., Langer, G. (2009). Synergies between transcription and lexical database building: The case of German Sign Language (DGS). In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of the Corpus Linguistics Conference (CL2009). University of Liverpool, UK, 20-23 July 2009*. [Online resource; URL: http://ucl.lancs.ac.uk/publications/cl2009/346_FullPaper.doc; last access: March 22, 2016].
- Kristoffersen, J.H. (2010). From utterance to example sentence – a lexicographic approach to exploitation of corpus. [Online resource; URL: http://www.tegnsprog.dk/hovedside/litteratur/SLCN_Berlin_2010_handout.pdf; last access: March 22, 2016].
- Kristoffersen, J.H., Troelsgård, T. (2010). Making a dictionary without words: lemmatization problems in a sign language dictionary. In S. Granger, M. Paquot (Eds.), *eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses universitaires de Louvain, pp. 165--172.
- Kristoffersen, J.H., Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language Dictionary. In S. Granger, M. Paquot (Eds.), *Electronic Lexicography*. Oxford: Oxford University Press, pp. 293--315.

- Langer, G., König, S., Matthes, S. (2014). Compiling a Basic Vocabulary for German Sign Language (DGS) – lexicographic issues with a focus on word senses. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus, July 15-19 2014 in Bolzano/Bozen – Italy*, pp. 767--786. [Online resource; URL: http://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/LangerKoenigMattesEURALEX2014.pdf; last access: March 22, 2016].
- Langer, G., König, S., Matthes, S., Groß, N., Hanke, T. (2016). Variation of DGS lexical items. What sign language lexicography can gain from a mixed method approach: Corpus data supplemented by crowd sourcing. Poster presented at the 12th International Conference on Theoretical Issues in Sign Language Research (TISLR12), Jan 4-7, 2016 at Melbourne, Australia. [Online resource; URL: http://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/TISLR_poster_mixedmethods_18k_korr_Version.pdf; last access: March 22, 2016].
- Svensén, B. (2009). *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.

9. Language Resource References

- Center for Tegnsprog (2008-2016). *Ordbog over Dansk Tegnsprog*. Available at: <http://www.tegnsprog.dk>.
- DGS Corpus Project (2015-2016): Subcorpus. Available at: <http://meine-dgs.de>.