Lemma Selection and Entry Types

Sabrina Wähl, Gabriele Langer, Anke Müller, Felicitas Otte, Thomas Hanke, Reiner Konrad

DOI (latest version): 10.25592/uhh.fdm10226

Version 0	2021-08-07	First draft
Version 1	2022-05-25	First version for public release

Zusammenfassung

Dieses Arbeitspapier ist Teil des lexikographischen Handbuchs für das Digitale Wörterbuch DGS (DW-DGS). Hier werden unsere Kriterien für die Lemmaselektion erläutert, der daraus resultierende Umfang des Wörterbuchs bestimmt und die verschiedenen Typen von Einträgen beschrieben. Außerdem werden Möglichkeiten für die Nutzung von vorverarbeiteten Daten diskutiert.

Abstract

These project notes are part of the lexicographic manual for the Digitales Wörterbuch DGS (DW-DGS). We will explain our lemma selection criteria, outline the resulting dictionary size, and describe the different entry types resulting from this process. Additionally, the possible uses of pre-processed data are discussed.

1 Corpus-based Lexicography of a Sign Language

A lemmatized corpus of a considerable size is the most important pre-condition for a detailed and reliable description of a signs' meanings and uses. Such a corpus allows for reviewing and analyzing a considerable amount of natural fluent signing in context. In modern lexicography, the standard procedure for Word Sense Discrimination (WSD) is to rely on corpus data.

For spoken languages in their written form, tools such as the Sketch Engine[1] support analysis of linguistic patterns by providing e.g. part of speech (POS) tagging, wordlists, and collocations. While this is standard for well-researched written languages such as English and German, it is largely not available for under-researched languages without a written tradition, script, and orthography including signed languages such as DGS. Thus, sign language lexicographers face specific challenges in preparing, reviewing, and processing a sufficient amount of data for the lexicographic analysis.

With the lack of automatic pre-processing and tagging and the time and effort that consequently go into manual annotation, such annotations are simply not available for sign languages on a large scale. For DGS – as for many other under-researched languages – there is an additional challenge: For the analysis of sentence structures and PoS tagging of corpus data, stabilized comprehensive and commonly agreed-upon grammatical descriptions and theories are needed. So far, there are none that could be applied robustly and consistently on a large scale. Hence even manual annotations on e.g. PoS prove to be a difficult task for DGS.

In addition, the review and summary of individual examples of sign use is more complex due to the video format – differing from a written representation by its fleetingness.

[1] https://www.sketchengine.eu

2 Lemma Selection and Lemma Establishment

2.1 DGS Corpus as Data Source: Annotations and Type Hierarchy

Our predominant data source for the compilation of dictionary entries is the DGS corpus[2] which consists of 91 hours of annotated data (06.08.2021). The corpus data is accessible for the lexicographic team through the annotational database and working environment of iLex. Here, lemmatization is done by matching tokens to a type-hierarchy containing types and subtypes connected to each other in a hierarchical structure on four levels depending on their form and meaning differences. In token-type-matching, a token can be assigned either to the general type (level 3) or one of its subtypes (levels 2,1,0) depending on the form and conventional meaning the individual token instantiates in context. Thus, the subtypes on the lower levels provide a rough indication of the form and meaning ranges a type can cover. Closer analysis of the data typically reveals a more fine-grained sense discrimination. For a detailed description of data organization, see Langer et al. 2016, and for the basic annotation see AP03-2018-01 (Konrad et al. 2020).

2.2 Lemma Sign Candidates

Lemma selection is generally driven by frequency in the corpus. In order to have enough data for a detailed analysis, we consider level-3 types as lemma candidates only if they have at least one subtype with a minimum of 25 tokens (cf. AP10-2016-02: Langer 2020). Applying this general threshold, our corpus contains 1674 lemma candidates (as of 2021-08).

2.3 Signs Included on the Basis of Additional Criteria and Data

In special cases, we have made an exception to the general threshold of 25 tokens, applied lower thresholds and included additional data sources, that is data obtained through the DGS Feedback (cf. Wähl et al. 2018) and SignHunter (cf. Hanke et al. 2020) for a better coverage of regionality and completeness of semantic sets or relevant names, e.g. city names.

Semantic sets with high regionality: There are certain groups of signs, such as names for months or days of the week, that vary greatly by region. As a result, signs that are used almost exclusively in smaller regions are only very weakly evidenced in the corpus, since only a few people from that specific region were filmed. If these regional signs were omitted in the dictionary due to the threshold of 25 tokens, certain regions might not be represented with their regional signs at all. Additionally, the representation of sign usage in the relevant meaning set, such as for months or days of the week, would be distorted. Therefore, we lowered the threshold of 25 occurrences in the corpus and additionally used data from the DGS Feedback[3] to get a clearer picture of the distribution of regional month and weekday signs.
In the DGS corpus, signs for the months were specifically elicited from 50% of all informants in the task ‘elicitation of isolated items’. Corpus evidence for these signs thus does not only consist of signs casually occurring in context in conversations but also of signs shown one after the other in a more artificial elicitation setting. Signs for days of the week were collected in a special task for finding appointments in a guided conversation. Through the DGS Feedback, additional data were obtained. The month and weekday signs found in the corpus were presented to participants who were asked to vote on whether they use the sign themselves, know the sign from other signers but do not use it themselves, or whether the sign was unknown to them.

[2] For more details about the data, see AP06-2020-01 (Schulder et al. 2021).
[3] The DGS Feedback System is an online survey system. For more details, see Wähl et al. 2018 or AP07-2014-01 (Matthes et al. 2014).

All signs for months and weekdays that were used by at least three people from the corpus and a total of at least five people (from the corpus and DGS Feedback combined) were included in the dictionary.

Other semantic sets displaying high regional variation e.g. color terms and kinship terms might also require similar treatment.

Regional lexical variants: Not only semantic sets but also some other concepts (e.g. 'to want', 'or', and 'healthy') are expressed in DGS with several lexical signs differing in their regional distribution. In some cases, there are only different regional lexical variants. Sometimes one predominant sign is used in large areas or all of Germany while competing smaller regional lexical variants exist. Following the same rationale as for the semantic sets with high regional variation, we aim to cover the regional distribution of the different signs also including smaller regional variants. As DGS Feedback data is not available for all of these concepts and signs, we set a threshold for additional, i.e. smaller regional variants to five DGS corpus occurrences from five different signers.

Essential synonyms and antonyms: When compiling a dictionary entry, one step is to provide information on signs with the same or similar meanings and signs with a contrasting meaning to the described sense. In the information structure of the entries, these synonyms or antonyms can support the differentiation and disambiguation of the respective sense. The synonyms and antonyms constitute elements of (cumulative) synonym definitions in DGS.[4] Some of these synonymous or antonymous signs that are especially valuable for this explanatory function may not be well evidenced in the corpus. If they are categorized as important for the understanding of the described sense, the synonym or antonym sign nevertheless may be included into the dictionary despite their thin evidence. The set threshold for these cases is 20 corpus tokens from at least seven participants.

City and country names: Another group of signs for which an exception of the 25 tokens rule was defined are city names and country names. The threshold for city names to be included in the dictionary is evidence from five participants in total, at least three of which must be participants of the DGS corpus. The others may be participants of the SignHunter elicitations. For country names, the threshold is five participants from the corpus.

2.4 Establishment of Lemma Signs / Lemmatization

At the beginning of the lexicographic process stands what we call lemma sign establishment (i.e. lemmatization in the lexicographic sense). During lemma sign establishment, we use editorial rules and data review processes to determine which types and subtypes of the type hierarchies in iLex constitute one dictionary entry, as well as which variants are included and shown in the entry.

Starting from the type hierarchies of the lemma sign candidates, the scope of the entries is determined. Some of the level-3 type candidates are variants of each other and are covered by the same entry. Other type hierarchies are split into several entries. It is also possible that a subtype is cut off to be described in its own entry while leaving the level-3 parent and sibling subtypes without enough tokens to qualify for another entry. Some level-3 types constitute entries that then additionally include other types as phonological variants that would not make it into entries for themselves[5].

[4] Dictionary definitions in the DW-DGS are given in German and not in DGS, therefore the listing of synonyms and antonyms are especially important from a monolingual perspective. For (cumulative) synonym definitions, see Svensén (2009, 214-217).

[5] For details see AP10-2016-01 (Langer et al. 2020).

2.5 Pre-Lemmatization

In 2021, all lemma sign candidates were pre-lemmatized. This resulted in 1608 pre-lemmatized entries, covering 2121 level-3 types and 2394 subtypes or qualified types altogether. Not included in this number are number sign roots and a few special cases. These are the entries that can be adequately described for the dictionary on the basis of the available corpus data. During further steps of the lexicographic process, some of the pre-lemmatization decisions may be corrected due to a more detailed look into the data or new data but we do not expect that the number of possible entries will change substantially.
Our aim is to provide entries for all 1600+ candidates in the dictionary. In order to reach this goal, we suggest a strategy of layered depth of coverage and detail for different entries. Our current pre-release entries in the DW-DGS vary in depth and size. Short and less detailed entries are e.g. given for those lemma signs that have only few occurrences but are included for the completion of regional coverage or semantic sets.

3 Different Entry Types

3.1 Detailed Entries

Lemma candidates show different behaviors in terms of semantic complexity and also differ in frequency of use. Quite often high frequency and semantic complexity go hand in hand. Detailed entries are entries that contain the full range of information of a sign as far as the data allows. With enough evidence at hand, a variety of information can be given in the entry:

- Sign form: main variant and additional variants
- Fine grained differentiation of meaning / senses
- Meaning and usage explanations
- Synonyms and antonyms
- Most frequent translational equivalents
- Indication of frequency
- Possible mouthings
- Regional distribution
- Cross-references concerning form
- Participation in multiword expressions (MWE [6]) such as collocations and compound-like constructions
- Examples taken from the corpus (usually two, though more can be given)

For a detailed description of the possible information types that may be given in the DW-DGS, see AP10-2021-02 (Langer et al. 2021).

Having a broad data base and thus the possibility of finding evidence for many senses of a sign, even those that are used less frequently, is both a blessing and a curse. On the one hand, a lexicographer may give a lot of information on a sign's uses that is potentially interesting for the future dictionary user. On the other hand, every lexicographer runs the risk of getting lost in the data and using up a lot of time by pursuing the lead of an interesting but less frequent meaning. Experience has shown that it is necessary and more efficient to focus on more frequent senses and to briefly document the minor ones in the dictionary database without preparing

[6] We use the term multiword expression (MWE) for multi sign units as the term is established in lexicography for all expressions that are composed of more than one linguistic unit.

them for display in the dictionary. This procedure is not uncommon in lexicography (cf. Atkins & Rundell 2008, 322-324).

For examples of detailed entries, see Entry 440 , Entry 80 , Entry 193. https://www.sign-lang.uni-hamburg.de/korpusdict/bags/bag193.html

3.2 Concise Entries

Concise entries are entries that are less elaborate with regard to the information that is given on the sign. This is due to the fact that the sign has a lower frequency in the corpus and thus less information can be given. The low frequency may influence the number and granularity of senses, as less frequent meanings may have too few observations to get described as individual senses in the dictionary. Instead, they may be lumped in with a closely related sense or they may not be presented in the dictionary at all. It also may affect information on collocations and the sign's participation in other MWEs if the number of tokens is not sufficient for the respective analysis. Additionally, it is possible that the usual number of example sentences is reduced due to the lack of suitable examples that can be drawn from the corpus.
Another factor that may play a role is the sign's semantic complexity. Some signs are used for a rather limited number of senses, e.g. Entry 629, which is used for three senses:

1) Protestant (denomination)
2) To pray
3) Church (as a building)

To sum up, concise entries differ from detailed entries with regard to:

- granularity and number of senses,
- possible lack of information on collocations and the sign's participation in MWEs,
- a possibly reduced number of examples sentences

Other examples of concise entries in the dictionary: Entry 459, Entry 150 , Entry 37

3.3 Short Entries

Short entries are entries with reduced information, i.e., they may contain fewer senses, the senses may have a rather rough granularity, and no information on collocations or the sign's participation in other MWEs is given. Example sentences are reduced to one example per sense or completely left out. No example is given in cases where the data does not contain a suitable example or when the meaning of the sense is so straightforward that examples are not considered essential for understanding.[7]

Short entries will be provided for signs with a less broad range of meaning, i.e., signs with one or few straightforward senses. We expect that many signs for objects, animals, cities, and colors will fall into this group. Another group of signs will probably consist of those regional signs that were elicited by means of the task 'elicitation of isolated signs' (cf. Nishio et al. 2010). If these signs are low frequency variants, they often only occur in that task, which makes an elaborate analysis difficult.

To sum up, short entries differ from detailed entries with regard to:

- granularity and number of senses

[7] Even when no example is selected for inclusion into the entry the sign can still be viewed in actual use when clicking on the concordance button below the entry. This provides a view on all occurrences of the sign in the Public DGS Corpus – with the movie just another click away.

- no information on collocations and the sign's participation in MWEs
- reduced number of example sentences (0-1 examples)

Examples of short entries: Entry 472 , Entry 535 , Entry 713

4 Pre-Processing

4.1 Automatically Generated Entries

In the pre-release entries (entries with red micon numbers) some cross-references point to pre-lemmatized entries (entries with white micon numbers, marked as *Automatisch generierter Vorabeintrag*). These automatically generated entries have either not yet been worked on or are being prepared but not finished and therefore not fed into the production by the lexicographic team. They are the product of pre-processing of information that can be drawn automatically from the corpus and the information in the already published dictionary entries (e.g. via cross-referencing). They are not displayed in any index but are given as the target of cross-referencing. Automatically generated entries will be replaced by edited entries as soon as they are ready for production.

Examples of automatically generated entries in the pre-release DW-DGS are: Entry 914, Entry 846, Entry 824 (as of 2022-05).

4.2 Pre-Processing of Entry Information

In the following phase, we will take pre-processing a step further and make use of pre-processed entry information as an additional view on the corpus data and evolving dictionary as much as possible in order to support and speed up the entry writing process. For this purpose, the lexicographic team will be provided with a view on pre-processed entry information (given outside the pre-release entries of the dictionary for entries that are in preparation). For lemma sign candidates of a low complexity of semantic range and low frequency – candidates for short entries – entry writing might basically consist of a review, selection or consolidation and correction of this information in order to assure the quality of the information presented in the entry, and adding definitions. For concise and detailed entries of lemma sign candidates with complex behavior and meaning ranges the pre-processed information can serve as a good starting point for the much more detailed analyses necessary.

Pre-processing is the automatic compilation of information using the data from the corpus and existing entries. The following information on a lemma sign could be automatically generated to support the lexicographic description of lemma signs:

information type	kind of information / function / gain	availability of information	additional work required or suggested
sign form and variants	defining scope of entry, sign is searchable and available for cross-referencing and indexing	already available through pre-lemmatization; HamNoSys notations for search for form function in iLex	production of missing movies

indication of frequency		number of occurrences in corpus	
mouthings	most frequent mouthings accompanying the sign [not sense-specific, for the whole entry]	mouthings are part of the corpus annotation	query and threshold; rough manual inspection and selection suggested
distributional maps	information of the regionality of the sign	can be generated from corpus data	decision whether all maps should be shown or whether a manual selection should take place
cross-references concerning form		hand-picked explicit cross-references from fully compiled entries; automatic references from iLex	for cross references between pre-processed entries an algorithm needs to be defined
participation in multiword expressions such as collocations and compound-like constructions	[not sense-specific, for the whole entry]	hand-picked explicit information transferred from fully compiled entries; automatic listing of frequent neighbors possible	for frequent neighbors in automatic entries an algorithm/threshold needs to be defined; a rough manual inspection and selection might be useful
indication of meaning / senses	indicated through synonyms and antonyms and their sense definitions	available when lemma sign is the target of cross-references from fully compiled entries	several cross-references produce multiple entries for the same sense that need to be lumped manually into one.
examples of use	concordance view of occurrences in the Public DGS Corpus		
most frequent translational equivalents	[not sense-specific, for the whole entry]	utterances with German translations	query that identifies the most frequent words in translations; manual inspection and selection strongly suggested

5 References

Atkins, B. T. Sue / Rundell, Michael (2008). *The Oxford guide to practical lexicography*. Oxford; New York: Oxford University Press.

Hanke, Thomas / Jahn, Elena / Wähl, Sabrina / Böse, Oliver / König, Lutz (2020). SignHunter – A Sign Elicitation Tool Suitable for Deaf Events. In Efthimiou, Eleni et al. (eds.): *Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages*. 12th International Conference on Language Resources and Evaluation (LREC 2020), Marseille, France, 2020. Paris, France: European Language Resources Association (ELRA), 83–88. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/20030.html.

Langer, Gabriele / Troelsgård, Thomas / Kristoffersen, Jette / Konrad, Reiner / Hanke, Thomas / König, Susanne (2016): Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing. In Efthimiou, Eleni et al. (Hrsg.): *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*. 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slowenien, 28.5.2016. Paris, Frankreich: European Language Resources Association (ELRA), 143–152. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/16014.html.

Nishio, Rie / Hong, Sung-Eun / König, Susanne / Konrad, Reiner / Langer, Gabriele / Hanke, Thomas / Rathmann, Christian (2010): Elicitation methods in the DGS (German Sign Language) Corpus Project. In Dreuw, Philippe et al. (Hrsg.): *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, 22.-23.5.2010. Paris, Frankreich: European Language Resources Association (ELRA), 178–185. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/10026.html.

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

Wähl, Sabrina / Langer, Gabriele / Müller, Anke (2018): Hand in Hand - Using Data from an Online Survey System to Support Lexicographic Work. In Bono, Mayumi et al. (Hrsg.): *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*. 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 12.5.2018. Paris, Frankreich: European Language Resources Association (ELRA), 199–206. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/18025.html.

6 Project Notes

Konrad, Reiner / Hanke, Thomas / Langer, Gabriele / König, Susanne / König, Lutz / Nishio, Rie / Regen, Anja. (2020). *Public DGS Corpus: Annotation Conventions / Öffentliches DGS-Korpus: Annotationskonventionen*. Project Note AP03-2018-01. Version 3. Hamburg, Germany: DGS-Korpus project, IDGS, Hamburg University. http://doi.org/10.25592/uhhfdm.1860.

Langer, Gabriele. (2020). *Vorgehen bei der Analyse für die Artikelschreibung (Wörterbuch).* Project Note AP10-2016-02. Version 4. Hamburg, Germany: DGS-Korpus project, IDGS, Hamburg University. http://doi.org/10.25592/uhhfdm.9577.

Langer, Gabriele / Müller, Anke / Wähl, Sabrina / König, Susanne / Hanke, Thomas / Konrad, Reiner. (2020). *Lemmatisierungsregeln (Vorläufige Wörterbuch-Einträge).* Project Note AP10-2016-01. Version 4. Hamburg, Germany: DGS-Korpus project, IDGS, Hamburg University. https://doi.org/10.25592/uhhfdm.1899.

Langer, Gabriele / Müller, Anke / Otte, Felicitas / Wähl, Sabrina. (2021). *Information Types and Use Cases*. Project Note AP10-2021-02. Version 1. Hamburg, Germany: DGS-Korpus project, IDGS, Hamburg University.

Matthes, Silke / Hanke, Thomas / König, Susanne / Langer, Gabriele. (2014). *Entwicklung des DGS Feedback-Systems*. Project Note AP07-2014-01. Version 1. Hamburg, Germany: DGS-Korpus project, IDGS, Hamburg University. http://doi.org/10.25592/uhhfdm.845.

Schulder, Marc / Blanck, Dolly / Hanke, Thomas / Hofmann, Ilona / Hong, Sung-Eun / Jeziorski, Olga / König, Lutz / König, Susanne / Konrad, Reiner / Langer, Gabriele / Nishio, Rie / Rathmann, Christian. (2021). *Data Statement for the Public DGS Corpus*. Project Note AP06-2020-01. Version 1. Hamburg, Germany: DGS-Korpus project, IDGS, Hamburg University. http://doi.org/10.25592/uhhfdm.9700.