Queries and Views in iLex to Support Corpus-based Lexicographic Work on German Sign Language (DGS)

Gabriele Langer, Anke Müller, Sabrina Wähl
Institute of German Sign Language and Communication of the Deaf, University of Hamburg
Gorch-Fock-Wall 7, 20354 Hamburg, Germany
E-mail: {gabriele.langer, anke.mueller, sabrina.waehl}@uni-hamburg.de

Abstract

In the DGS-Korpus project the corpus is being used as the basis for lexicographic descriptions of signs in dictionary entries. In this process the lexicographers start from the data and type entry structures as found in the annotation database. While preparing a dictionary entry much of the work consists of manually going through a number of single tokens viewing the original data and available annotations. Findings are then categorised and summarised. However, a number of decisions and descriptions are also supported by pre-defined searches and views on the data. Supported areas include lexicographic lemmatisation (lemma sign establishment), selection of citation forms and variants, grammatical behaviour of signs, collocational patterns of use, regional distribution patterns and distribution of lexical or formational variants over different age groups. While we are still in the process of exploring the possibilities of a sign language corpus for lexicography, searches and views that have proven useful for our work are exemplified in this paper with regard to dictionary entries.

Keywords: corpus-based lexicography, corpus searches, German Sign Language (DGS)

1. Introduction

One of the central aims of the DGS-Korpus project is the compilation of a corpus-based dictionary of German Sign Language (DGS). The basis for a lexicographic description of signs is the reference corpus that was collected within the project for lexicographic and other purposes (Blanck et al., 2010). Corpus data is accessed through the annotational and lexical database and working environment iLex (Hanke & Storz, 2008). During the lexicographic work on preparing an entry it is essential that the available data can be viewed easily and quickly. While working on different entries similar basic analytical questions regarding a sign's properties re-occur with regard to different signs. It is helpful that such questions can be answered quickly through pre-defined queries and views on the data (cf. Atkins & Rundell, 2008: 103, 104).

2. Corpus-based Dictionary of DGS

The dictionary of DGS being produced within the DGS-Korpus project is the first corpus-based dictionary of DGS. Its aim is a description of signs and their use as found in the corpus. Lexicographic descriptions and decisions are informed directly by the analysis of the available corpus data. Dictionary entries include example sentences for the described sign senses directly taken from the originally recorded corpus material.

3. DGS-Korpus Data

The corpus of the DGS-Korpus project was intended to serve as basis for the dictionary from the very beginning. Some elicitation tasks (*Subject Areas, Calender Task, Regional Specialities* and *Elicitation of Isolated Signs*) were specifically included for this purpose (cf. Nishio et al., 2010). The data consists of signed conversations, narrations, discussions, retellings, and other sign uses of 330 informants filmed between 2010 and 2012. Informants from all over Germany were included and balanced for gender, four age groups and 13 regions. For the balancing of regions the estimated population size of sign users was taken into account. Informants were filmed in pairs in one-day sessions. Nearly 560 hours of signing were recorded, up to now 64 hours are completed for basic lemmatisation and annotation. Lemmatisation and annotation is ongoing.[1] Material that is not yet or will not be lemmatized is to a large degree at least translated and can be searched via the translations for specific concepts. In some cases this leads to spot annotations of relevant passages. The corpus size is now approx. 465.000 tokens (23.02.2018).

4. iLex

iLex is the annotational and lexical database and working environment that is used in the DGS-Korpus project for annotating corpus data. It is – up to now – also the only tool that we use to access and view the DGS-Korpus data for the purpose of a lexicographic description of signs.[2]
In iLex, type entries are created to represent abstract sign types to which occurrences of signs (i.e. tokens) are linked. Two type entries can be related to each other in superordinate-subordinate relationship: each type can have only one superordinate type, while a superordinate type may have a number of subordinate types. The user of iLex can define the number of type levels they need in order to set up their data structures.
iLex also provides the user with the possibility to define, store, and re-use SQL-queries and to generate

[1] Lemmatisation here is token-type-matching and an important part of the basic annotation. Lemmatisation in the lexicographic sense may follow different criteria to decide on which elements are attributed lemma sign status and receive their own dictionary entry. To avoid confusion, lemmatisation in the lexicographic sense will here be called *lemma sign establishment* following a suggestion of Svensén (2009: 94).

[2] Other ways to access the data are described and discussed in Jahn et al. (2018) also in this issue.

distributional maps and other visualisations directly from the data (Hanke, 2016).[3]

5. Annotational Type Structures

It is helpful to know how the corpus data is structured in our annotational database in order to better understand the views shown in this paper. In the DGS-Korpus project we use two main and two secondary type levels to build a hierarchical *type structure*[4] that pre-structures the token evidence belonging to one sign.

A sign – an abstract independent meaningful unit of DGS – with all its forms and meanings is represented by a type entry at the highest level, called *sign* in the iLex type structure. A sign is defined and distinguished from other signs by its abstract form, overall range of meaning, and by its underlying image, in cases where its form has been iconically motivated. A *sign* type entry is represented in iLex by a unique gloss[5] and a specific citation form noted in HamNoSys. Instantiations (i.e. tokens) of a sign usually can be identified as belonging to a specific conventional use or meaning of this sign. Such established uses of a sign are modelled in our iLex database as subtypes called *lexemes. Lexemes* are subtype entries that are subordinate to *sign* entries. They group tokens that share one of the conventional meanings of the sign. *Lexeme* entries are specified by a unique gloss, a HamNoSys noting their citation form, and a rough indication of their conventional meaning through the assignment of concepts. Each *lexeme* belongs to exactly one *sign* while one (polysemous) *sign* can have a number of *lexemes* attached to it. Tokens that belong to the sign but cannot or have not yet been identified as established uses are not matched at the *lexeme* level but on the *sign* level within the type structure.[6]

The two main type levels of *signs* and *lexemes* are used in the basic annotational lemmatisation process, the token-type-matching. In a second step we also want to gain an impression of the different realisations of the form a sign can take – be it formational (or phonological) variation, grammatical or iconic modification or simply the range of realisations due to performance factors. For that purpose tokens differing from the citation form of *signs* or *lexemes* are grouped by adding recurring form features to the *sign* or *lexeme* gloss. These features name the difference to the citation form by the way of descriptive categories with feature values that are added to the *sign* or *lexeme* gloss. The categories are called *qualifiers* and the resulting groupings are called *qualified signs* or *qualified lexemes*. In iLex, these groupings are modelled as type entries subordinate to *signs* or *lexemes.* Their form is described by HamNoSys notation.[7] Tokens connected to a *qualified sign* or *qualified lexeme* are instantiations of the corresponding superordinate type or subtype.

In the views *sign* glosses are marked by an additional -$SAM at the end to indicate glosses of the highest type level (e.g. TIME1-$SAM). Lexical variants and non-related signs that share the same gloss word are distinguished by numbers (e.g. OR1 vs. OR2). Formational variants are distinguished from each other by letters following the number (e.g. OLD2A vs. OLD2B). *Sign, lexeme*, *qualified sign,* or *qualified lexeme* type entries are created in the lexical database only when needed for annotation. Thus, the hierarchical type structure belonging to one *sign* and the pre-sorting of tokens through that hierarchy provide a first structured view on the corpus data for the respective *sign* (cf. fig. 2 and fig. 8 in apx.).

6. Preparing dictionary entries

A dictionary entry aims at describing the typical uses of words – or in our case signs – disregarding rather untypical uses in order to inform the addressee of the dictionary about how to understand or use a respective item (e.g. Atkins & Rundell 2008: 54, 272). To this aim the lexicographer interprets, weighs and summarizes corpus findings and sometimes other sources of information in a user-oriented, standardised way.

Preparing a corpus-based dictionary entry involves a number of different steps. Atkins and Rundell (2008: 98-103) describe the first stage of this process as the *analysis* stage. The lexicographer reviews and analyses the available data and stores all noticeable facts about the sign in a pre-dictionary database which will serve at a later stage – the *synthesis* stage – as the basis for writing the actual dictionary entry.

In the DGS-Korpus project we are now at the analysis stage of preparing entries based on corpus data.[8] For this, it is essential that all data concerning a sign can be

[3] Self-written queries can be located at different spots within iLex. For example *display filters* define the information to be displayed in lists of items such as type lists; *lists* define the contents to be seen in tabs of display windows, e.g. a subtype list for a supertype or a token list in the type window.

[4] Terms we use to refer to entities and elements in our iLex database are indicated by italics.

[5] iLex uses a relational database. Token-type-matching is internally done via automatically generated IDs. Therefore glosses do not need to bear the function of IDs (as the ID-glosses in Johnston, 2008). They are nothing but unique labels for sign types for the practical handling while working with the data. A gloss can be easily changed without any effect on the lemmatisation and results. In the way we set up our structures in iLex, a constraint prohibits that two different types can be given the same gloss. The new gloss will appear in all transcripts, type entries and views automatically. For the purpose of making the DGS-Korpus publicly accessible for an international audience, types have also been given English glosses. In this paper, we have changed most views to display English glosses instead of German ones.

[6] Each token belonging to a *lexeme* at the same time also belongs to the superordinate *sign* and thus can be identified by both glosses – the *lexeme* gloss or the *sign* gloss, depending on which level of abstraction one wants to focus. This is what we call *double glossing* (cf. Konrad et al., 2012).

[7] As annotation is an ongoing process, *qualifiers* have been defined and introduced to iLex for a number of recurring form features corresponding to modification and variation kinds, but not for all occurring ones.

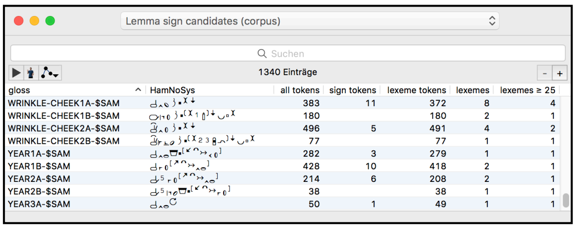
[8] This paper focuses on corpus data. We also use data obtained by an online survey system on signs and their use called the DGS-Feedback. For how we use data from the DGS-Feedback see Wähl et al. (2018) in this issue.

accessed easily from the corpus. We store our findings in a FileMaker database which at the moment serves as our pre-dictionary database. This database usually contains more information on a sign than what will appear in an actual dictionary entry. Elements of the proto-entry in this pre-dictionary database are marked for publication. Preliminary entries are then produced from exports of this database converted by scripts into an html structure. Representative studio recordings of single signs and original corpus examples prepared in iLex for publication are added to the preliminary entries.

In the remainder of the paper, we will discuss different queries and views in iLex that we have created and found helpful for analysis and decision-making when working on dictionaries entries. We will do so by roughly following the different steps of the workflow. Examples are given to show how corpus data can help answer questions that are relevant to lexicographic decisions and descriptions of signs. Our topic is not to discuss how to construct SQL-queries but what kind of views and pre-stored queries have proven useful in the process.

6.1 Lemma Sign Selection

Sign types are taken as lemma sign candidates and frequency counts help to estimate which types have enough data to enter the lexicographic process. Figure 1 shows a *filter* displaying a type list with a frequency count of attached tokens and the number of subtypes (*lexemes*) with a token count of 25 and above. We consider 25 tokens a minimum number necessary for a description of sign senses of a conventional use of a sign (i.e. a *lexeme*).[9]


Lemma sign candidates (corpus)

1340 Einträge

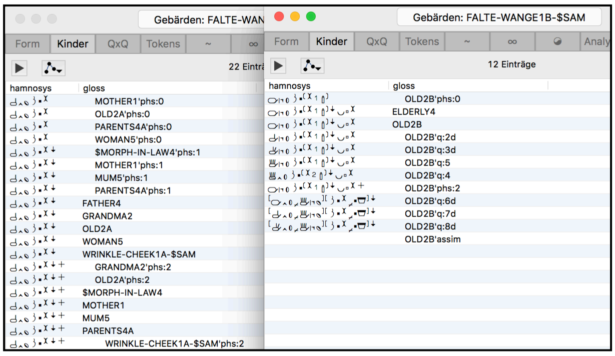
gloss	HamNoSys	all tokens	sign tokens	lexeme tokens	lexemes	lexemes ≥ 25
WRINKLE-CHEEK1A-$SAM		383	11	372	8	4
WRINKLE-CHEEK1B-$SAM		180		180	2	1
WRINKLE-CHEEK2A-$SAM		496	5	491	4	2
WRINKLE-CHEEK2B-$SAM		77		77	1	1
YEAR1A-$SAM		282	3	279	1	1
YEAR1B-$SAM		428	10	418	2	1
YEAR2A-$SAM		214	6	208	2	1
YEAR2B-$SAM		38		38	1	1
YEAR3A-$SAM		50	1	49	1	1

Figure 1: Part of lemma sign candidate list

6.2 Establishment of Lemma Signs

While establishing a lemma sign many different aspects have to be considered. The starting point for an entry is the corpus evidence as it presents itself in the pre-structured way of the annotational database. First, the lexicographer needs to decide, according to the lemmatisation rules of the dictionary, what portion of the data is best described together in one entry or where to split the data into more than one entry. Lexicographic decisions can follow different rules than annotational decisions and may result in a partly different grouping of evidence (cf. Langer et al., 2016). Lemma sign establishment requires an overview of the type structure of the lemma sign candidate and possibly related signs (variants similar signs). The list view of the lemma sign candidates (see fig. 1) already gives an impression of related signs. The type structure of these signs can be displayed and compared within list views showing the *lexemes* and their *qualified forms* (see fig. 2).

The *signs* a) WRINKLE-CHEEK1A-$SAM and b) WRINKLE-CHEEK1B-$SAM are formationally and iconically related. With respect to those characteristics, they might be phonological variants and constitute one single lemma sign. The signs both show *lexemes* with the meaning[10] 'old' but only sign a) can also mean e.g. 'woman', 'mother' or 'grandma'. Additionally, only the *lexeme* of sign b) with the meaning 'old' can undergo numeral incorporation[11]. So, difference in evidenced meanings and grammatical behaviour are two good reasons to describe the signs as two different lemma signs and thus in two entries. Nevertheless cross-references between the dictionary entries will be made because of their iconic and formational relationship. Thus, dictionary users can easily find similar and related signs.


Gebärden: FALTE-WAN

22 Einträ

hamnosys	gloss
	MOTHER1'phs:0
	OLD2A'phs:0
	PARENTS4A'phs:0
	WOMAN5'phs:0
	$MORPH-IN-LAW4'phs:1
	MOTHER1'phs:1
	MUM5'phs:1
	PARENTS4A'phs:1
	FATHER4
	GRANDMA2
	OLD2A
	WOMAN5
	WRINKLE-CHEEK1A-$SAM
	GRANDMA2'phs:2
	OLD2A'phs:2
	$MORPH-IN-LAW4
	MOTHER1
	MUM5
	PARENTS4A
	WRINKLE-CHEEK1A-$SAM'phs:2

Gebärden: FALTE-WANGE1B-$SAM

12 Einträge

hamnosys	gloss
	OLD2B'phs:0
	ELDERLY4
	OLD2B
	OLD2B'q:2d
	OLD2B'q:3d
	OLD2B'q:5
	OLD2B'q:4
	OLD2B'phs:2
	OLD2B'q:6d
	OLD2B'q:7d
	OLD2B'q:8d
	OLD2B'assim

Figure 2: (*Qualified*) *lexemes* of the *signs* a) on the left and b) on the right

6.3 Main Variant and Citation Form

The data to be described in one particular dictionary entry may contain several different sign forms – be it form variants, morphologically relevant modifications or just differences due to performance. For example phonetic variance such as one-handed vs. two-handed occurrences or non-morphological variance in movement repetition can be observed. In a dictionary entry, one form is chosen to represent the whole lemma sign in all its occurring forms. This form is called *lemma* or *citation form*. The lexicographer needs to decide which variants to display in the entry, which variant to choose as main variant and which form of this variant to choose as citation form. Summarised listings of occurring sign forms with token counts are available for a description of form variants in the dictionary (see fig. 3).

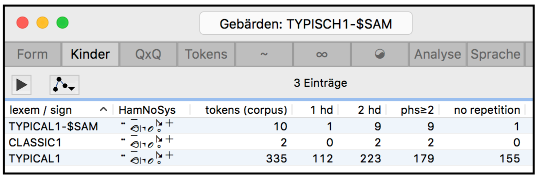
Criteria for the choice of main variant can be a higher frequency, broader regional distribution, and broader range of meaning. The corpus data can help to decide what the main variant might be. A query sorting out the

[9] This is a somewhat arbitrary number we chose relating to Sinclair (2005: 11) who suggests a minimum of at least 20 instances necessary for an outline description of the behaviour of a not particularly ambiguous word. Depending on the properties of the respective *lexeme* more evidence might be necessary.

[10] At this stage Word Sense Disambiguation (WSD) still has to be conducted. Thus the given meanings are preliminary and do not specify the whole range of senses the signs may have in the dictionary entry.

[11] This morphological difference is marked by the *qualifier q:* and the corresponding number being incorporated (see fig. 2). The letter *d* behind the number signifies that the handshape includes the thumb.

frequency of forms can be executed, meaning-related as well as form-related. The *sign* TYPICAL1-$SAM is phonologically simple and exhibits some phonetic variation with respect to handedness and repetition. Figure 3 shows an overall distribution of these features and gives an impression of the most frequent forms.



lexem / sign	HamNoSys	tokens (corpus)	1 hd	2 hd	phs≥2	no repetition
TYPICAL1-$SAM		10	1	9	9	1
CLASSIC1		2	0	2	2	0
TYPICAL1		335	112	223	179	155

Figure 3: Summary for number of hands and repetition

The main variant of the lemma sign TYPICAL1-$SAM seems to be two-handed. As for repetition, the picture is not straight forward; but taking into account the many contextual or performative uses of one-handed forms in general, the main variant tends to include repetition. Having summaries of evidence for occurring form variations helps the lexicographer to make an informed decision on a citation form for the entry.

6.4 Description of Meaning (WSD)

The core task for the lexicographer is a documentation of the evidenced range of meaning. This is described by the way of dictionary *senses*. Usually this entails looking through corpus data by the way of a KWIC[12] view on the data – that is a selection of concordance lines (Atkins & Rundell, 2008: 311). The lexicographer groups the contextual meanings of the tokens and describes them as senses and sub-senses in the pre-dictionary database (a process also called Word Sense Disambiguation (WSD), cf. Atkins & Rundell 2008: 269).

In iLex, a number of different views on the data are available when working on analysing, categorising, and summarising the meaning range of a sign. While preparing a sign entry, many tokens are reviewed one by one in context. The analyser views both original recording as well as the corresponding annotations. When the token numbers do not allow the analysis for all tokens in detail, the most promising ones are selected. A token list displaying *lexeme* and *qualified lexeme* glosses, mouthings, translations, left and right neighbours, informants, region, and data collection tasks supports making an informed choice – covering a variety of people, regions, subjects and linguistic context (see fig. 4, apx.).

In iLex the view corresponding to a KWIC list is called *tokens in context*. This view is implemented in iLex and can be filled as needed by the iLex user through suitable queries. For selected tokens, it provides important information such as a sign string, mouthings, translation. Additional information on informant, region and elicitation task is displayed in the lower part for the activated line (see fig. 5, apx.).

Since the DGS data are not written in nature, they do not allow for quick browsing. Available annotations can support but not fully replace viewing the original movie. The original recording corresponding to the selected *tokens-in-context* line can be opened and viewed quickly.

Another view we find helpful for the WSD is the view of frequent left and right neighbours that is described in the next chapter. Collocational patterns can help to identify different uses of a sign with regard to meaning (cf. Atkins & Rundell 2008: 301-304; Kilgarriff, 2012: 7).

6.5 Collocational Patterns

Co-occurrence patterns not only help to distinguish sign senses, but they are also used to identify collocational patterns, idiomatic phrases, and compounds or compound-like combinations. This is supported by a view listing frequent left and right neighbours of the *lexemes* of a given *sign* type (see figure 6, apx.).[13] The view lists neighbours of a lexeme when this combination appears at least five times in the corpus. The view also shows the mutual information score and the number of pattern tokens and informants for that pattern.

Figure 6 (see apx.) shows the co-occurrence results for the type structure of TIME1-$SAM (ZEIT1-$SAM). Marked in blue are combinations that can be interpreted as compound-like sign strings shadowing the elements of German compounds usually accompanying the signs in form of respective mouthings, e.g. YEAR TIME1 (German compound *Jahres|zeit*) or TIME1 PRESSURE (German compound *Zeit|druck*). Often, these compound-like patterns are not fixed combinations of two particular lemma signs but more dynamic combinations. For example, there are three different lexemes YEAR1, YEAR2 and YEAR3 (with in total six formational variants) contributing to the pattern YEAR TIME1.

Marked in red are combinations with number signs or number-incorporating signs used for indicating the time of the day. Marked in orange are other combinations also relating to the time of the day. Two central meanings of the sign can be identified through the green combinations 'good time' and 'beautiful time' versus the yellow combinations NONE/MORE/MUCH-OR-MANY TIME1, TO-NEED TIME1, TIME1 BARELY, and TIME1 FOR. The green combinations are typical of the sense that can be described as 'a specific period in history or a person's life'. The yellow combinations are typical for the following sense of 'time': 'a resource that is needed or available to conduct some activity and that can be plentiful, limited, scarce or lacking.' Lists of frequent neighbours can also indicate typical arguments or argument groups of predicate signs and possibly it will also be helpful to detect idiomatic phrasal structures. In lexicography, we use the list of frequent neighbours

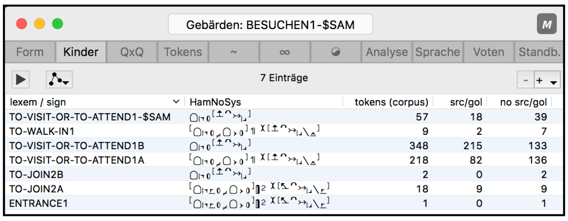
[12] KWIC = keyword in context.

[13] This view on the corpus data is based on a formula that has been used in the Sketch Engine up to September 2006 to determine the mutual information score (MI) (see Lexical Computing, 2015: 2). In our annotational database sometimes even small formational differences are differentiated by new type or subtype entries in order to be able to detect and analyse regional differences. One result of this is smaller token numbers for each grouping. For the purpose of WSD, such finer-grained distinctions are conflated into more general groupings in the co-occurrence list. This is done in a rather coarse way by leaving off additional numbers and letters behind the gloss names (which normally indicate lexical and formational variants) when running the co-occurrence analysis (cf. fig. 6, apx.). The analysis is sensitive with regard to meaning and therefore is run on the *lexeme* level. Individual neighbouring subtypes contributing to the pattern are listed in the last column of the view.

especially for WSD. Dictionary entries are planned to include typical collocational patterns and compound-like combinations, which are selected from the neighbours' list.

6.6 Grammatical Behaviour

In the annotation process, some of the annotated *qualifiers* refer to or serve as a possible indicator of grammatical properties of a sign, such as numeral incorporation or spatial behaviour. There are list views that serve to find candidates for grammatical behaviour and show the existence or presumed non-existence of selected *qualifier* features. One view summarises all tokens with one particular feature regardless of the individual values. *Qualifiers* directly signalling grammatical behaviour are for example *source/goal* or *goal*, expressing form features of indicating or so-called agreement verbs (see fig. 7 for the *sign* TO-VISIT-OR-TO-ATTEND1-$SAM); or *body location*, referring to the morphological change of place of articulation with respect to body parts, or *qualifiers* noting number incorporation (cf. fig. 2). Indicators of possibly grammatical behaviour are *qualifiers* for *phases* (i.e. repetition) as they can point to aspect forms, as well as alterations of speed and size.



lexem / sign	HamNoSys	tokens (corpus)	src/gol	no src/gol
TO-VISIT-OR-TO-ATTEND1-$SAM	[illegible]	57	18	39
TO-WALK-IN1	[illegible]	9	2	7
TO-VISIT-OR-TO-ATTEND1B	[illegible]	348	215	133
TO-VISIT-OR-TO-ATTEND1A	[illegible]	218	82	136
TO-JOIN2B	[illegible]	2	0	2
TO-JOIN2A	[illegible]	18	9	9
ENTRANCE1	[illegible]	1	0	1

Figure 7: Summary for feature *source/goal*

For the interpretation of the figures it is important to consider the stage of the annotation process (adding *qualifiers* is part of the detailed annotation at a later stage). This means that not all modified forms (*src/gol*) may already be marked. Also, as the form from centre to front is the citation form in the annotational database, only those tokens receive a *qualifier* whose forms differ from the citation form.

For a closer look, other list views show the total range of token forms of a given type and its subtypes, giving the count of tokens for each individual form (see fig. 8, apx.). This view is helpful to gain an overview of the sign's possible characteristics (and how they are distributed with respect to different senses as roughly represented by subtypes), but also to pick out interesting cases. This way usage restrictions or subsenses connected to certain formational behaviour can be found. Since annotation and lemma revision are ongoing processes, the presented situation is not yet fully fledged but helps to detect grammatical behaviour and larger formational sign classes.

6.7 Regional Distribution

A dictionary description of the regional distribution of a lemma sign is easily supported by the rendering of maps as visual representations of distributional patterns (Hanke et al., 2017). For the lexicographic work we regularly use two kinds of maps that show either token numbers or numbers of informants using the sign(s), that is types or subtypes, in question. The grading of regionality follows our data collection subregions within Germany. Maps can easily be rendered directly from the data in iLex by marking the respective types or subtypes in a type list and selecting the desired pre-stored map kind.

The first map kind visualizes the use of the selected type or subtype by indicating the number of tokens (or, if desired, informants) by a colouring from white to yellow to orange to dark red in eight steps. This map gives a good impression on where the sign or lexeme is used and where the core areas of use are. See for example the number of informants using the *lexeme* OR3 in figure 9 (see apx.).

The second kind of maps visualises and contrasts the use of a cluster of lexical and formational variants for presumably the same concept. For each subregion the number of informants using the types or subtypes (or if selected: number of tokens used in that region) are displayed as a pie chart. The pie charts' size is relative to the total number of items and regions are coloured with the colour of the item with the strongest evidence. See for examples the variant cluster for the *lexemes* with the meaning 'or' (fig. 10, apx.).

The map kind 2 (cluster of *lexemes*) shows regional differences and confirms that the regional distribution as shown in the map kind 1 is not the result of still missing data from other regions but truly a result of the use of different variants.

6.8 Age Related Sign Use (Language Change)

Language change is another aspect to be considered while writing an entry, as information on age groups and their preference of signs or sign variants is valuable information on sign use. Signs that show less and less usage along age groups descending from "senior" to "junior" may be prone to vanish and therefore are marked in the entry as "dated". This can occur with respect to specific meanings of a sign (as represented by *lexemes*), or to all meanings. In the latter case the whole sign would be regarded as dated. To detect patterns of language change, clusters of *lexemes* of the same meaning can be compared with respect to the four age groups established.[14] It is advisable to look at clusters and not only isolated lexemes, to minimise effects of chance distribution and get more reliable results (cf. Hanke et al., 2017). For example, the *lexemes* TO-MOVE2 and TO-MOVE1 from different *sign* types are both used to denote 'to move (change of residence)'. The two signs differ in handshape and show a considerable age effect, which we can see via doughnut charts that visualise age distribution with possible clusters. The count can either be on tokens or on different informants. Informant count is more significant here. Two different views have proven helpful. Fig. 11 shows the distribution of informants from the four age groups per *lexeme*. A balanced overall distribution of informants on age groups (with respect to the signs compared) is a prerequisite for a reliable result, which is met in this example (see fig. 11, apx., doughnut on the right).

[14] Based on a date of reference (01.01.2011) the corpus informants were grouped into age groups. The years of birth of the defined age groups lie between 1981-94 for the defined age group 18-30, 1966-1980 for the age group 31-45, 1965-1951 for the age group 46-60 and ≤ 1950 for the age group 61+. People from the cohort ≥1995 have not been included in the corpus because they were not of age at the time of recording.

Another type of doughnut chart view highlights the number of informants of a certain age group using TO-MOVE2 or TO-MOVE1 (see fig. 12 in apx.). Here, the increase of use of TO-MOVE1 can be seen from left to right, where the left doughnut represents the oldest informants and the right one the youngest. With those instruments to analyse the use of signs with respect to age groups, possible trends can be discovered and documented.

7. Conclusion

Corpus-based lexicography of a sign language is a comparatively new field as larger corpora of these non-written languages are now becoming available. Not all of the tools and methods developed for written languages can be directly or effortlessly applied to sign corpora. However, even today, corpus data can already answer many questions on sign use more reliably than it was possible before. The process of developing and experimenting with useful ways to annotate, analyse, summarise and visualise sign corpus data for the needs of sign lexicography is ongoing, and we continuously improve and add to our queries and views on the data. From our experience, we are convinced that in the future sign lexicography will benefit even more from corpora when annotation conventions and analysis methods are further developed.

8. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

9. Bibliographical References

Atkins, B.T.S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Blanck, D., Hanke, T., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., König, L., König, S., Konrad, R., Langer, G., Nishio, R., Rathmann, C., Vorwerk, S., Wagner, S. (2010). The DGS Corpus Project. Development of a Corpus Based Electronic Dictionary German Sign Language – German. Poster presented at the Theoretical Issues in Sign Language Research Conference (TISLR 10), Sept 30 - Oct 2, 2010 at Purdue University, Indiana, USA.

Hanke, T., Konrad, R., Langer, G., Müller, A., Wähl, S. (2017). Detecting Regional and Age Variation in a Growing Corpus of DGS. Poster presented at the 9th International Corpus Linguistics Conference (CL 2017) pre-conference workshop "Corpus-based Approaches to Sign Language Linguistics: Into the Second Decade", Jul 24, 2017 at Birmingham, UK.

Hanke, T. (2016). Towards a Visual Sign Language Corpus Linguistics. In E. Efthimiou, E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, & J. Mesch (Eds.), *Corpus Mining. Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages. 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.* Paris: ELRA, pp. 89--92.

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. Zwitserlood, & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages. 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco*. Paris: ELRA, pp. 64--67.

Jahn, E., Konrad, R., Langer, G., Wagner, S., Hanke, T. (2018). DGS-Korpus Project: Different Formats for Different Needs. In this issue.

Johnston, T. (2008). Corpus linguistics and signed languages: No lemmata, no corpus. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages. 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco*. Paris: ELRA, pp. 82--87. Paris: European Language Resources Association.

Kilgariff, A. (2012). Using Corpora [and the Web] as Data Sources for Dictionaries. Online resource; URL: https://www.sketchengine.co.uk/wp-content/uploads/Using_corpora_2012.pdf [Accessed 2018-02-19].

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. *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.

Langer, G., Troelsgård, T., Kristoffersen, J., Konrad, R., Hanke, T., König, S. (2016). Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing. In E. Efthimiou, E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, & J. Mesch (Eds.), *Corpus Mining. Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages. 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.* Paris: ELRA, pp. 143--152.

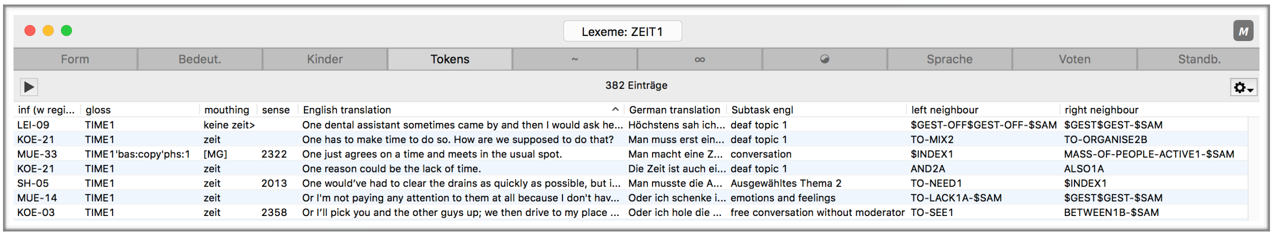
Lexical Computing Ltd. (2015): Statistics used in the Sketch Engine. July 8, 2015. Online resource; URL: http://www.sketchengine.co.uk/wp-content/uploads/ske-statistics.pdf [Accessed 2017-01-18].

Sinclair, J. (2005). Corpus and Text – Basic Principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxford Books, 1--16. Available online from: http://ota.ox.ac.uk/documents/creating/dlc/ [Accessed 2018-02-19].

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

Wähl, S, Langer, G., Müller, A. (2018). Hand in Hand – Using Data from an Online Survey System to Support Lexicographic Work. In this issue.

10. Appendix

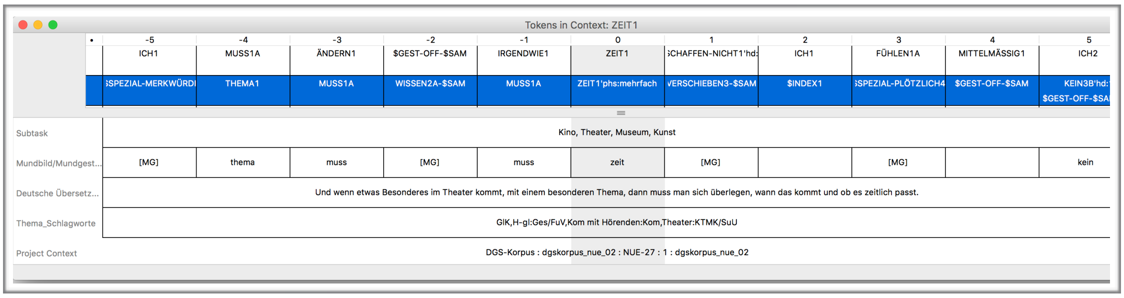


Lexeme: ZEIT1

382 Einträge

inf (w regi...	gloss	mouthing	sense	English translation	German translation	Subtask engl	left neighbour	right neighbour
LEI-09	TIME1	keine zeit>		One dental assistant sometimes came by and then I would ask he...	Höchstens sah ich...	deaf topic 1	$GEST-OFF$GEST-OFF-$SAM	$GEST$GEST-$SAM
KOE-21	TIME1	zeit		One has to make time to do so. How are we supposed to do that?	Man muss erst ein...	deaf topic 1	TO-MIX2	TO-ORGANISE2B
MUE-33	TIME1'bas:copy'phs:1	[MG]	2322	One just agrees on a time and meets in the usual spot.	Man macht eine Z...	conversation	$INDEX1	MASS-OF-PEOPLE-ACTIVE1-$SAM
KOE-21	TIME1	zeit		One reason could be the lack of time.	Die Zeit ist auch ei...	deaf topic 1	AND2A	ALSO1A
SH-05	TIME1	zeit	2013	One would've had to clear the drains as quickly as possible, but i...	Man musste die A...	Ausgewähltes Thema 2	TO-NEED1	$INDEX1
MUE-14	TIME1	zeit		Or I'm not paying any attention to them at all because I don't hav...	Oder ich schenke i...	emotions and feelings	TO-LACK1A-$SAM	$GEST$GEST-$SAM
KOE-03	TIME1	zeit	2358	Or I'll pick you and the other guys up; we then drive to my place ...	Oder ich hole die ...	free conversation without moderator	TO-SEE1	BETWEEN1B-$SAM

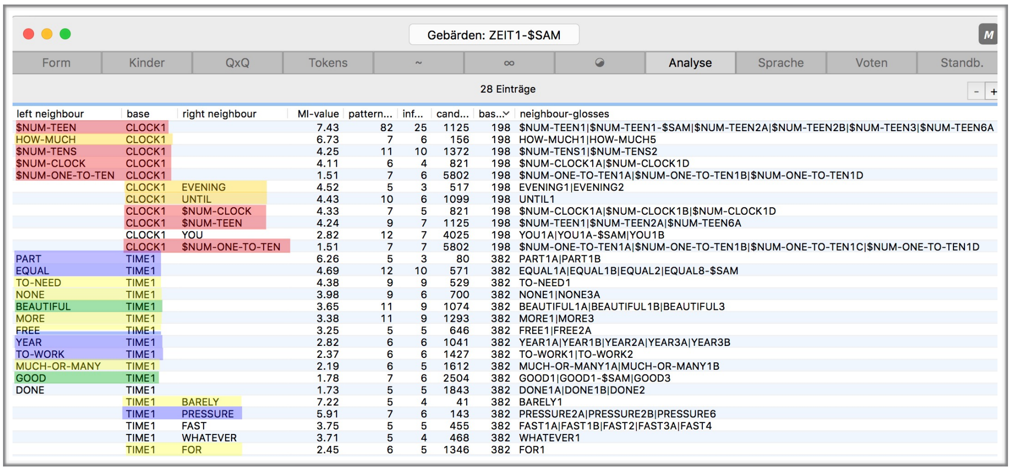
Figure 4: Token list of the *lexeme* TIME1 with relevant information for WSD



Tokens in Context: ZEIT1

•	-5	-4	-3	-2	-1	0	1	2	3	4	5
	ICH1	MUSS1A	ÄNDERN1	$GEST-OFF-$SAM	IRGENDWIE1	ZEIT1	;CHAFFEN-NICHT1'hd:	ICH1	FÜHLEN1A	MITTELMÄSSIG1	ICH2
	;SPEZIAL-MERKWÜRDI	THEMA1	MUSS1A	WISSEN2A-$SAM	MUSS1A	ZEIT1'phs:mehrfach	VERSCHIEBEN3-$SAM	$INDEX1	;SPEZIAL-PLÖTZLICH4	$GEST-OFF-$SAM	KEIN3B'hd: $GEST-OFF-$SA
Subtask	Kino, Theater, Museum, Kunst										
Mundbild/Mundgest...	[MG]	thema	muss	[MG]	muss	zeit	[MG]		[MG]		kein
Deutsche Übersetz...	Und wenn etwas Besonderes im Theater kommt, mit einem besonderen Thema, dann muss man sich überlegen, wann das kommt und ob es zeitlich passt.										
Thema_Schlagworte	GIK,H-gl:Ges/FuV,Kom mit Hörenden:Kom,Theater:KTMK/SuU										
Project Context	DGS-Korpus : dgskorpus_nue_02 : NUE-27 : 1 : dgskorpus_nue_02										

Figure 5 : View *tokens in context*

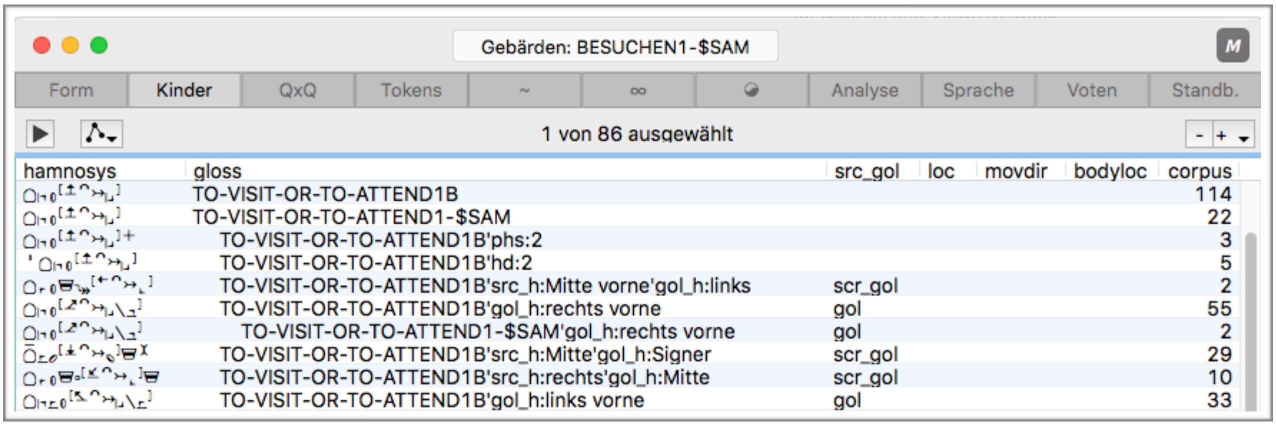


Gebärden: ZEIT1-$SAM

28 Einträge

left neighbour	base	right neighbour	MI-value	pattern...	inf...	cand...	bas...	neighbour-glosses
$NUM-TEEN	CLOCK1		7.43	82	25	1125	198	$NUM-TEEN1\|$NUM-TEEN1-$SAM\|$NUM-TEEN2A|$NUM-TEEN2B\|$NUM-TEEN3|$NUM-TEEN6A
HOW-MUCH	CLOCK1		6.73	7	6	156	198	HOW-MUCH1|HOW-MUCH5
$NUM-TENS	CLOCK1		4.25	11	10	1372	198	$NUM-TENS1\|$NUM-TENS2
$NUM-CLOCK	CLOCK1		4.11	6	4	821	198	$NUM-CLOCK1A\|$NUM-CLOCK1D
$NUM-ONE-TO-TEN	CLOCK1		1.51	7	6	5802	198	$NUM-ONE-TO-TEN1A\|$NUM-ONE-TO-TEN1B|$NUM-ONE-TO-TEN1D
	CLOCK1	EVENING	4.52	5	3	517	198	EVENING1|EVENING2
	CLOCK1	UNTIL	4.43	10	6	1099	198	UNTIL1
	CLOCK1	$NUM-CLOCK	4.33	7	5	821	198	$NUM-CLOCK1A\|$NUM-CLOCK1B|$NUM-CLOCK1D
	CLOCK1	$NUM-TEEN	4.24	9	7	1125	198	$NUM-TEEN1\|$NUM-TEEN2A|$NUM-TEEN6A
	CLOCK1	YOU	2.82	12	7	4025	198	YOU1A|YOU1A-$SAM|YOU1B
	CLOCK1	$NUM-ONE-TO-TEN	1.51	7	7	5802	198	$NUM-ONE-TO-TEN1A\|$NUM-ONE-TO-TEN1B|$NUM-ONE-TO-TEN1C\|$NUM-ONE-TO-TEN1D
PART	TIME1		6.26	5	3	80	382	PART1A|PART1B
EQUAL	TIME1		4.69	12	10	571	382	EQUAL1A|EQUAL1B|EQUAL2|EQUAL8-$SAM
TO-NEED	TIME1		4.38	9	9	529	382	TO-NEED1
NONE	TIME1		3.98	9	6	700	382	NONE1|NONE3A
BEAUTIFUL	TIME1		3.65	11	9	1074	382	BEAUTIFUL1A|BEAUTIFUL1B|BEAUTIFUL3
MORE	TIME1		3.38	11	9	1293	382	MORE1|MORE3
FREE	TIME1		3.25	5	5	646	382	FREE1|FREE2A
YEAR	TIME1		2.82	6	6	1041	382	YEAR1A|YEAR1B|YEAR2A|YEAR3A|YEAR3B
TO-WORK	TIME1		2.37	6	6	1427	382	TO-WORK1|TO-WORK2
MUCH-OR-MANY	TIME1		2.19	6	5	1612	382	MUCH-OR-MANY1A|MUCH-OR-MANY1B
GOOD	TIME1		1.78	7	6	2504	382	GOOD1|GOOD1-$SAM|GOOD3
DONE	TIME1		1.73	5	5	1843	382	DONE1A|DONE1B|DONE2
	TIME1	BARELY	7.22	5	4	41	382	BARELY1
	TIME1	PRESSURE	5.91	7	6	143	382	PRESSURE2A|PRESSURE2B|PRESSURE6
	TIME1	FAST	3.75	5	5	455	382	FAST1A|FAST1B|FAST2|FAST3A|FAST4
	TIME1	WHATEVER	3.71	5	4	468	382	WHATEVER1
	TIME1	FOR	2.45	6	5	1346	382	FOR1

Figure 6: Frequent left and right neighbours of the *sign* TIME1-$SAM



Gebärden: BESUCHEN1-$SAM

1 von 86 ausgewählt

hamnosys	gloss	src_gol	loc	movdir	bodyloc	corpus
	TO-VISIT-OR-TO-ATTEND1B					114
	TO-VISIT-OR-TO-ATTEND1-$SAM					22
	TO-VISIT-OR-TO-ATTEND1B'phs:2					3
	TO-VISIT-OR-TO-ATTEND1B'hd:2					5
	TO-VISIT-OR-TO-ATTEND1B'src_h:Mitte vorne'gol_h:links	scr_gol				2
	TO-VISIT-OR-TO-ATTEND1B'gol_h:rechts vorne	gol				55
	TO-VISIT-OR-TO-ATTEND1-$SAM'gol_h:rechts vorne	gol				2
	TO-VISIT-OR-TO-ATTEND1B'src_h:Mitte'gol_h:Signer	scr_gol				29
	TO-VISIT-OR-TO-ATTEND1B'src_h:rechts'gol_h:Mitte	scr_gol				10
	TO-VISIT-OR-TO-ATTEND1B'gol_h:links vorne	gol				33

Figure 8: Summary of sign forms of TO-VISIT-OR-TO-ATTEND-$SAM with token counts (segment)

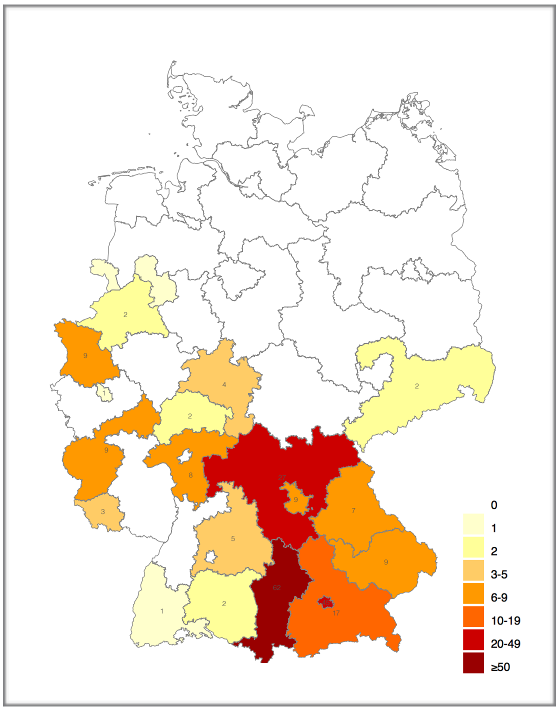


Figure 9: Map (tokens) for *lexeme* OR3

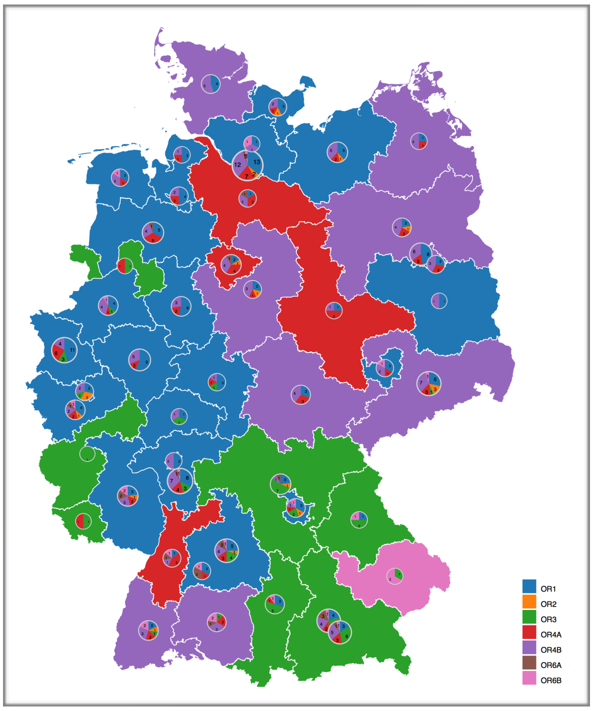


Figure 10: Map (informants) for variant cluster „or“

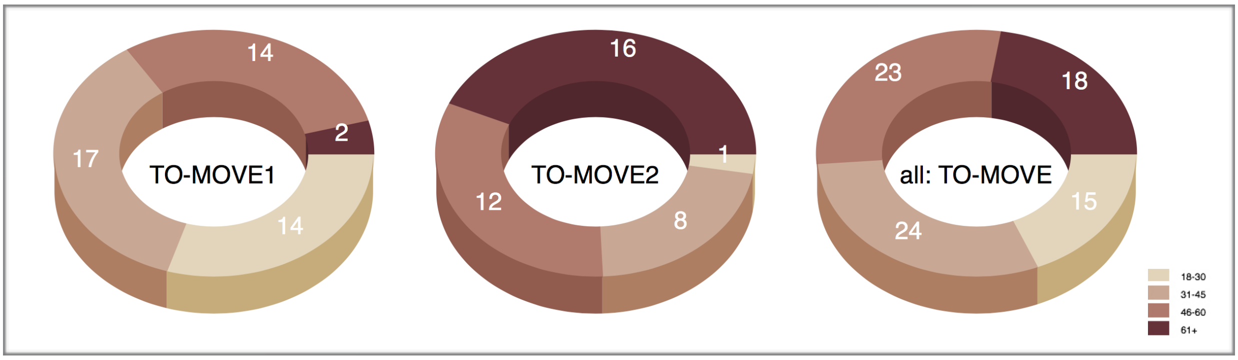


Figure 11: Doughnut charts (informants' age groups per *lexeme*) for „to move“

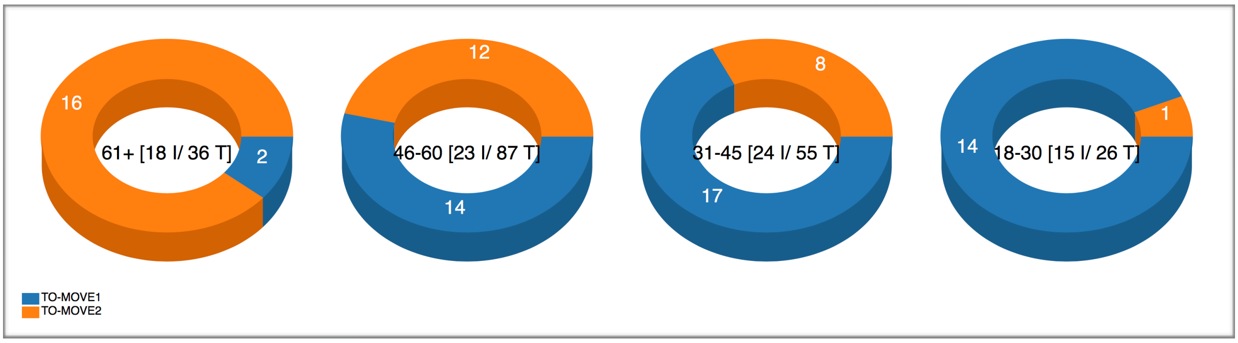


Figure 12: Doughnut charts (*lexemes* per informants' age group)