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

Gabriele Langer, Anke Müller, Sabrina Wähl
University of Hamburg, Institute for German Sign Language and Communication of the Deaf, Germany

Background: DGS-Korpus

- Project: DGS-Korpus (2009-2023)
- Goals: reference corpus & corpus-based dictionary
- Tool: iLex
- Notation system: HamNoSys (for type forms)
- Data:
  - 330 informants
  - Balanced for 13 regions, 4 age groups and gender
  - Filmed in pairs at 12 locations (2010-2012)
  - Data collection tasks included signed conversations, narrations, discussions, and retellings (cf. Nishio et al. 2010)
- DGS Corpus (26.4.2018):
  - Nearly 560 hrs of signing (raw data)
  - Approx. 352 hrs translated into German and time-aligned on task level or utterance level (approx. 71 hrs), ongoing
  - Approx. 64 hrs completely lemmatised, ongoing
  - Approx. 480.000 tokens (including spot annotations)

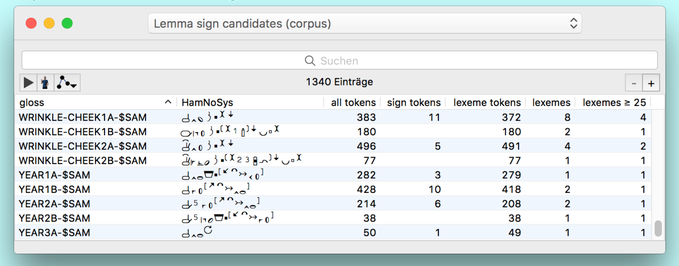
iLex

- Database and annotation environment (cf. Hanke 2002)
- Videos and time-aligned annotations
- Inventory of types (type entries) for lemmatisation
- iLex uses internal numerical type IDs. This frees glosses from having to serve as IDs.
- iLex structures can be modeled to needs of individual project:
  - Number and function of tiers
  - Hierarchical structures of types
  - SQL queries (can be stored in iLex for re-use)
  - Maps and other visualisation formats generated directly from the data (Hanke 2016)

1. Lemma Sign Selection

Lexicographic question: Which signs should be described in the dictionary?

- Listing of *signs* with corpus evidence (frequency count of tokens for *lexemes* ≥ 25) indicates lemma sign candidates for lexicographic analysis and processing.



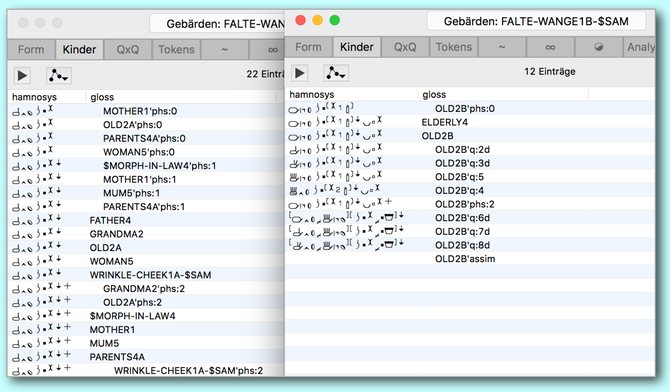
gloss	HamNoSys	all tokens	sign tokens	lexeme tokens	lexemes	lexemes ≥ 25
WRINKLE-CHEEK1A-$SAM		383	11	372	8	4
WRINKLE-CHEEK1B-$SAM		180		180	2	1
WRINKLE-CHEEK2A-$SAM		496	5	491	4	2
WRINKLE-CHEEK2B-$SAM		77		77	1	1
YEAR1A-$SAM		282	3	279	1	1
YEAR1B-$SAM		428	10	418	2	1
YEAR2A-$SAM		214	6	208	2	1
YEAR2B-$SAM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

Lemma sign candidate list with token counts

2. Lemma Sign Establishment (Lemmatisation)

Lexicographic question: Which sections of the corpus data (type hierarchies) should be presented in one entry and which parts are better presented in separate entries? (cf. Langer et al. 2016)

- Type hierarchies pre-structure data for analysis (e.g. subtype list with occurring qualified forms). Data of variant candidates can be compared. The data of the candidates below suggest separate entries (due to differences in meaning and modification behaviour).

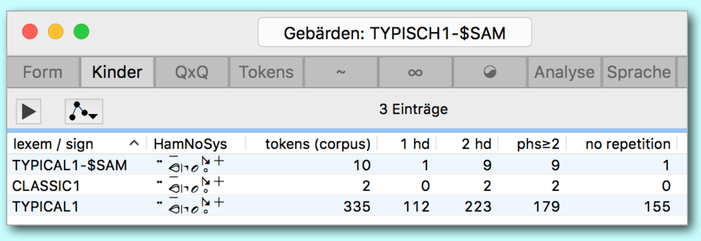


Type entries showing details of two variant candidates

3. Main Variant and Citation Form

Lexicographic questions: Which formational variants of a sign exist? Which variant should be chosen as the main variant? What sign form should represent the lemma sign as citation form in the dictionary?

- Criteria for the choice of the main variant are higher frequency, broader regional distribution, and broader range of meaning.
- Summarised listings of occurring sign forms show the frequency of form variants, e.g. phonetic variation in the number of hands and repetition and thus support the decision on the citation form.



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

Summary for number of hands and repetition for TYPICAL1-$SAM

- Frequency counts suggest the two-handed and repeated form as citation form for TYPICAL1-$SAM.

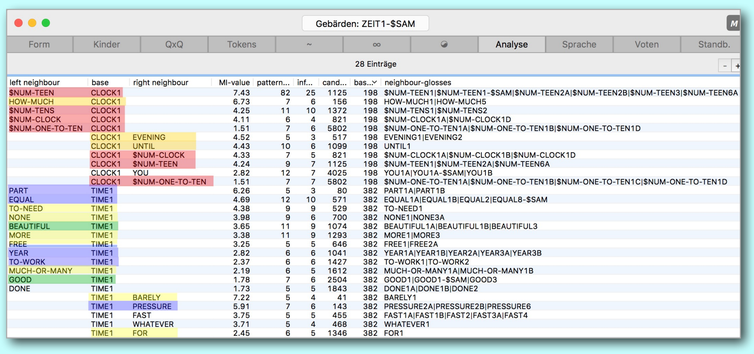
Sign Language Lexicography

- Corpus-based lexicography: DGS signs and their use are described in the dictionary on the basis of available corpus data.
- The type hierarchy of the annotational database (iLex) pre-structures available corpus data.
- Similar basic analytical questions regarding a sign's properties re-occur with regard to different lemma sign candidates.
- Useful queries, views, and visualisations are pre-stored in our iLex database to support lexicographic analysis, decision-making, and description at various stages of the process.
- In the dictionary entry, findings are summarised in the description of a sign's properties.

4. Collocational Patterns

Lexicographic question: What are typical left and right neighbours of the sign?

- MI (mutual information) score is used to identify frequent combinations (cf. Lexical Computing Ltd., 2015).
- Co-occurrence patterns indicate collocational patterns, compound-like combinations, and possibly idiomatic phrases.
- Collocational patterns can help to identify and distinguish sign senses.
- Typical patterns and combinations are included in the dictionary entry.



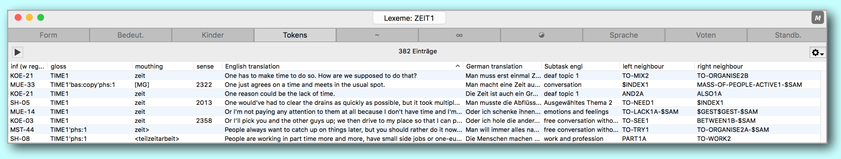
left neighbour	base	right neighbour	MI-value	pattern...	inf...	cand...	bas...	neighbour-glosses
$NUM-TEEN	CLOCK1		7.43	82	25	1125	198	$NUM-TEEN1\|$NUM-TEEN1-$SAM\|$NUM-TEEN2A|$NUM-TEEN2B\|$NUM-TEEN3|$NUM-TEEN6A
HOW-MUCH	CLOCK1		6.73	7	6	156	198	HOW-MUCH1|HOW-MUCH5
$NUM-TENS	CLOCK1		4.25	11	10	1372	198	$NUM-TENS1\|$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

Frequent left and right neighbours of the sign TIME1-$SAM

5. Word Sense Disambiguation (WSD)

Lexicographic question: What are the typical uses/meanings of the sign?

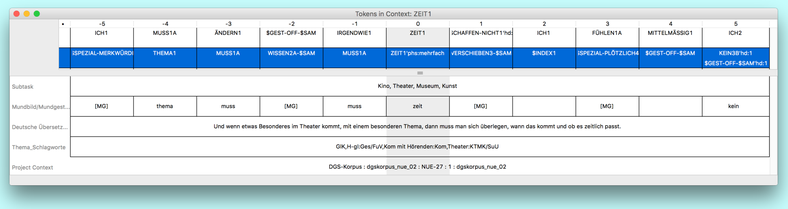
- Examination of many tokens in context (taking different regions, persons, situations etc. into account)
- Pre-selection supported by token lists displaying relevant information (on region, informant, formational properties via qualifiers, mouthing, translation, data collection task, left and right neighbours)



lnf (w reg...	gloss	mouthing	sense	English translation	German translation	Subtask engl	left neighbour	right neighbour
KOE-21	TIME1	zeit		One has to make time to do so. How are we supposed to do that?	Man muss erst einmal Z...	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 Zeit au...	conversation	$INDEX1	MASS-OF-PEOPLE-ACTIVE1-$SAM
KOE-21	TIME1	zeit		One reason could be the lack of time.	Die Zeit ist auch ein Gr...	deaf topic 1	AND2A	ALSO1A
SH-05	TIME1	zeit	2013	One would've had to clear the drains as quickly as possible, but it took multipl...	Man musste die Abflüss...	Ausgewähltes Thema 2	TO-NEED1	$INDEX1
MUE-14	TIME1	zeit		Or I'm not paying any attention to them at all because I don't have time and I'm...	Oder ich schenke ihnen ...	emotions and feelings	TO-LACK1A-$SAM	$GEST$GEST-$SAM
KOE-03	TIME1	zeit	2358	Or I'll pick you and the other guys up; we then drive to my place so that I can p...	Oder ich hole die ander...	free conversation witho...	TO-SEE1	BETWEEN1B-$SAM
MST-44	TIME1'phs:1	zeit>		People always want to catch up on things later, but you should rather do it now...	Man will immer alles na...	free conversation witho...	TO-TRY1	TO-ORGANISE2A-$SAM
SH-08	TIME1'phs:1	<teilzeitarbeit>		People are working in part time more and more, have small side jobs or one-eu...	Die Menschen machen ...	work and profession	PART1A	TO-WORK2

Token list of *lexeme* TIME1 with relevant information for WSD

- *Tokens-in-context* view (similar to KWIC lists, concordance views)
  - Allows to efficiently browse through and examine a list of selected tokens (see upper part of window) and their immediate linguistic context via gloss string representations of segments
  - Displays further annotations (mouthings, translation) and relevant data (metadata) for selected line (see lower part of window)
  - Quick access to movie corresponding to segment (original DGS data)

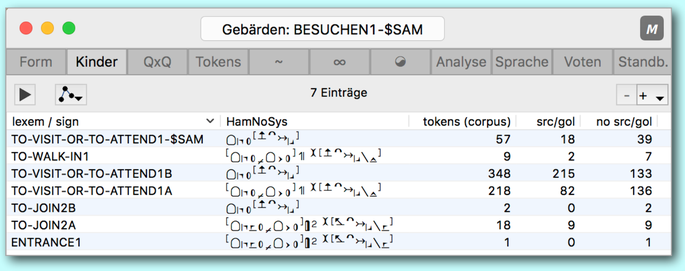


Tokens-in-context View

6. Grammatical Behaviour

Lexicographic question: What are the grammatical properties of the lemma sign? What kind of regular modifications occur?

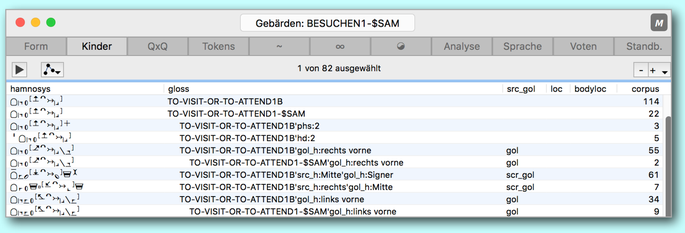
- List views for specific *qualifiers* summarise the occurrence of certain form features across the *lexemes* of a *sign*, thus suggesting a sign type as e.g. indicating verb being modified for source and goal.



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

Summary for feature *source/goal* with token counts

- A second kind of list view shows the distribution of various feature values as evidenced in the corpus, for all *qualifiers* realised.

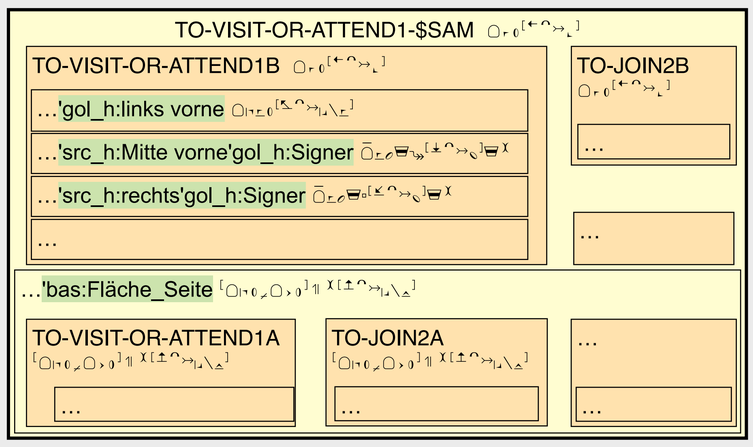


hamnosys	gloss	src_gol	loc	bodyloc	corpus
	TO-VISIT-OR-TO-ATTEND1B				114
	TO-VISIT-OR-TO-ATTEND1-$SAM				22
	TO-VISIT-OR-TO-ATTEND1B'phs:2				3
	TO-VISIT-OR-TO-ATTEND1B'hd:2				5
	TO-VISIT-OR-TO-ATTEND1B'gol_h:rechts vorne	gol			55
	TO-VISIT-OR-TO-ATTEND1-$SAM'gol_h:rechts vorne	gol			2
	TO-VISIT-OR-TO-ATTEND1B'src_h:Mitte'gol_h:Signer	scr_gol			61
	TO-VISIT-OR-TO-ATTEND1B'src_h:rechts'gol_h:Mitte	scr_gol			7
	TO-VISIT-OR-TO-ATTEND1B'gol_h:links vorne	gol			34
	TO-VISIT-OR-TO-ATTEND1-$SAM'gol_h:links vorne	gol			9

Summary of *sign* forms of a type with token counts

DGS Corpus Data Structures for Annotation in iLex

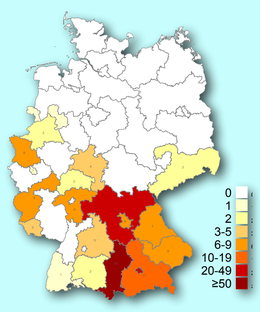
- Type hierarchies with four levels for modeling relevant differences in iconicity, form, and use/meaning of a sign (cf. Konrad et al. 2012)
- Two main type levels:
  - Types (*signs*) – represent signs as abstract linguistic units; in iLex: with a HamNoSys citation form; gloss ending with: -$SAM
  - Subtypes *(lexemes)* – represent established uses of a sign with regard to variant forms and meaning (pre-sorting loosely by meanings); in iLex: with a HamNoSys citation form
- Each *lexeme* belongs to exactly one *sign;* a *sign* may have several *lexemes* attached
- Tokens of a conventional use linked to corresponding *lexeme*; other tokens linked to *sign*
- Two secondary type levels: *qualified signs* and *qualified lexemes*
  - Recurrent form differences to citation form further grouped by adding descriptive categories (*qualifiers)* and *values* to *sign* or *lexeme* gloss
  - To mark formational (or phonological) variation, grammatical or iconic modification, or range of realisations due to performance factors



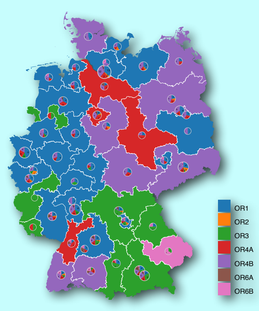
7. Regional Variation

Lexicographic questions: Where is the lemma sign used? Where are its formational or lexical variants used?

- Maps for number of tokens (or informants) can be generated directly from corpus data supporting the analysis of regional variation of signs.
- Left map shows tokens of OR3 indicating core areas of use.
- Right map contrasts number of informants using different lexical and formational variants of the *lexeme* cluster for 'or'.



OR3 'or': 225 tokens (of 49 informants); date: 2018-02-20

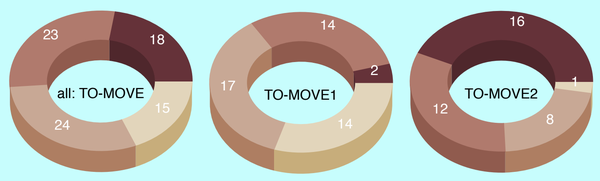


Variant cluster OR 'or': 248 informants (with 1365 tokens); date: 2018-02-20

8. Age Related Use (Language Change)

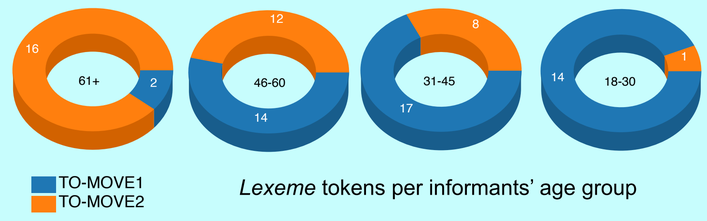
Lexicographic question: Is the lemma sign used mainly by older or younger signers? Is it becoming dated?

- Doughnut charts visualise age distribution of *lexeme* clusters, comparing the use of different signs with the same meaning.
- The use of the concept TO-MOVE ('to change residence') is evenly spread over all age groups. The form TO-MOVE1 is mainly used by younger, while the form TO-MOVE2 is used by older persons.



Number of informants for each age group per *lexeme*

- The chart below shows the progression of possible language change (apparent time).



Lexeme tokens per informants' age group

References

Hanke, T. (2002). iLex. A tool for Sign Language Lexicography and Corpus Analysis. In M. González Rodriguez, & 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: www.lrec-conf.org/proceedings/lrec2002/pdf/330.pdf [Accessed: 2018-04-27].

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.

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: 2018-04-27].

Nishio, R., Hong, S.-E., König, S., Konrad, R., Langer, G., Hanke, T. & Rathmann, C. (2010). Elicitation methods in the DGS (German Sign Language) Corpus Project. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz, & A. Schembri (Eds.), *Corpora and Sign Language Technologies. Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages. 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta. Paris: ELRA, pp. 178--185.

Poster presented at the *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. LREC 2018.* Miyazaki, Japan. May 12th, 2018.

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.