The DGS Corpus Project

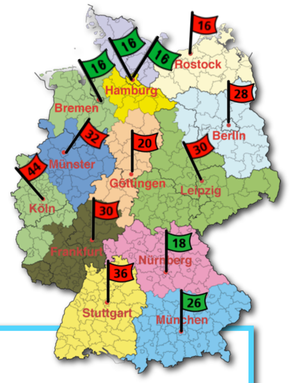
Development of a Corpus Based Electronic Dictionary *German Sign Language – German*

Dolly Blanck, Thomas Hanke, Ilona Hofmann, Sung-Eun Hong, Olga Jeziorski, Thimo Kleyboldt, Lutz König, Susanne König, Reiner Konrad, Gabriele Langer, Rie Nishio, Christian Rathmann, Stephanie Vorwerk, Sven Wagner

University of Hamburg, Institute of German Sign Language and Communication of the Deaf

Background Information

- Duration: 15 years (2009-2023)
- Responsible organization: Academy of Sciences in Hamburg, Germany
- Realization: Institute of German Sign Language and Communication of the Deaf (University of Hamburg)
- Funding: € 8.4 millions of the German Academies of Science Programme and additional resources provided by the University of Hamburg
- 150 person-years of work
- Team: 9 researchers, 4 Deaf research assistants, 1 technical staff, up to 18 student research assistants
- Goals:
  - Reference corpus of German Sign Language (DGS)
  - Corpus-based Dictionary of DGS - German



Involvement of the Deaf Community

- 328 informants
- 22 regional contact persons (find informants, raise public awareness on the project within the language community in their region)
- Focus group (planned): approximately 20-25 deaf experts (regionally rooted representatives of language community) provide support and validate analyses
- Feedback: web-based portal (planned) to receive feedback from members of language community e.g. on usage and regional variation of lexical items
- Deaf team members, researchers and student co-workers
- Flow of information between project and community through presentations at local Deaf clubs, focus group, website, facebook etc.

Transcription and Annotation

Translation

- Translation into German, segmentation into utterances

Basic Transcription

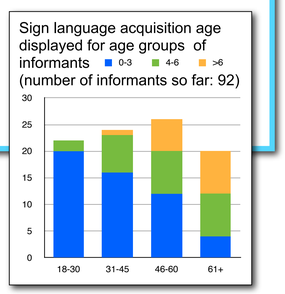
- Transcription / annotation carried out by student research assistants
- Supervised and checked by native signers
- Tokenization (segmentation into single signs)
- Lemmatization (token-type matching: identification and tagging of lexical items via glossing), tagging of productive signs and other signs
- Further specifications:
  - Variant, modified and deviant sign forms
  - Mouthings

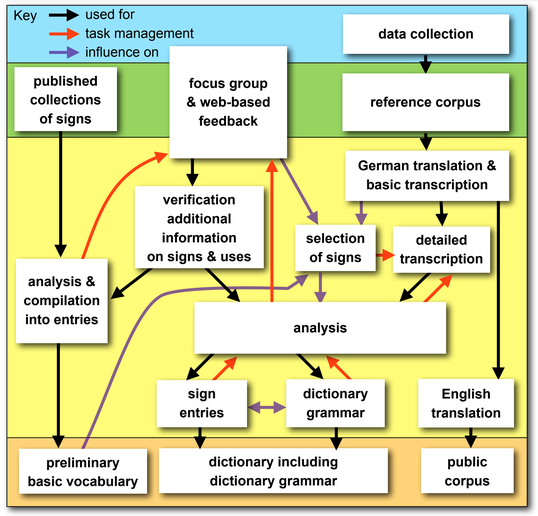
Detailed Transcription

- Approximately 50% of the basic transcriptions will be transcribed in more depth as needed for analysis and dictionary production
- Differentiation of phonological variants, grammatical sign forms (e.g. plural, negation, modifications), use of space
- Coding of contextual meaning
- Syntactic categories
- Sign context
- Mouth gesture, (lexical) facial expressions
- Sub-sentence phrase structure

Data Collection

- 2009-2011
- 13 regions at 12 locations (mobile studio)
- 328 informants balanced for sex, age, region
- Studio session:
  - 2 informants (peer-to-peer situation)
  - Moderated by Deaf contact person
  - Duration: one day (approximately 5.5 hours of filming plus breaks)
- Elicitation tasks (about 20 different tasks)
  - Various stimuli (e.g. signed texts, movies, pictures, words)
  - Different subjects areas to cover basic vocabulary
  - Various text types: e.g. conversation, discussion, description, re-telling, planning



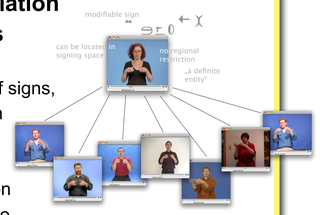


The iLex Environment

- Transcription and annotation tool
- Works with several synchronized video streams allowing the user to switch between different perspectives
- Integrated lexical database supports token-type-matching
- Metadata integrated into the database
- Multi-user approach
- Analyses via SQL statements
- Support of lexicographic workflow (work in progress)
- Support of quality assurance (work in progress)
- Export functions to ELAN, Quicktime with subtitles, HTML etc.
- Integrates video processing

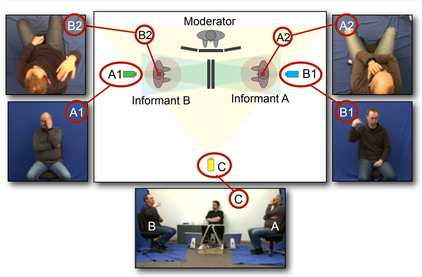
Analysis and Compilation of Dictionary Entries

- Analysis of spatial and grammatical behaviour of signs, contextual meaning, form variation, usage
- Abstraction from corpus data and other information (feedback, focus group) to give a general description of lexical signs, their forms, meanings and uses, variation, dialectal information



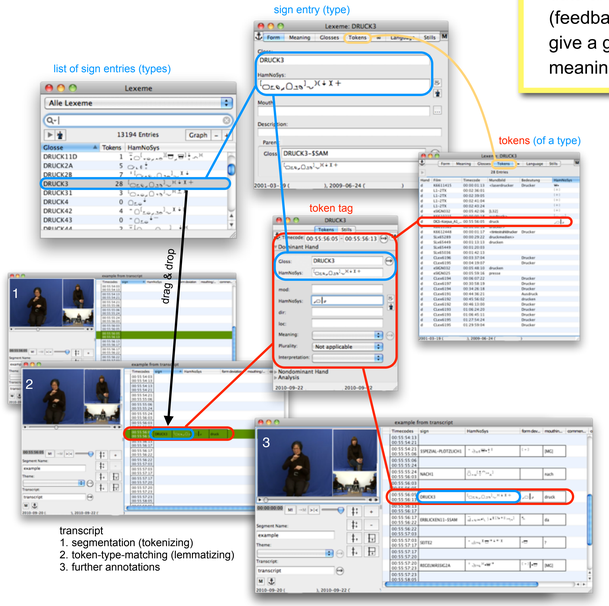
Studio Setup

- Mobile studio
- 7 cameras for 5 recording perspectives
- A1 and B1: front views on informants: HD cameras & stereoscopic cameras
- A2 and B2: birds-eye views on informants: HD cameras
- C: whole scene: HD camera
- 12 computers



The Reference Corpus

- 328 informants: men and women, 4 age groups, 13 regions
- 350-400 hours of footage
- 2.25 million tokens (estimated)
- 500 TB raw data (expected)
- Metadata on informants' linguistic and social background and studio session (IMDI standard)
- Tokenized, lemmatized and annotated
- Uses: basis for dictionary entries, language documentation, resource for basic linguistic research, resource for Deaf studies (texts on Deaf experiences and lives, Deaf culture), signed texts usable for sign language teaching



The Dictionary

- Corpus-based
- Descriptive
- In electronic form
- 6000 sign entries (planned)
- Bidirectional: search via sign form or written word
- Sign entries including information on form, meaning, grammar, variants and usage
- Examples of use taken from the corpus
- Cross references to related and similar signs
- Dictionary grammar
- We are currently experimenting with search by sample function.
- To be published in 2023

Preliminary Basic Vocabulary

- Basic vocabulary of DGS and German
- Based on evaluation of published sign collections (not on corpus data)
- Signs verified by focus group and web-based public feedback
- To be published in 2013
- To be replaced by the general dictionary in 2023

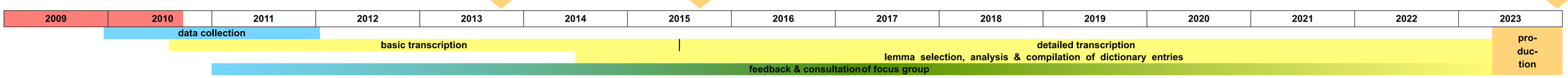
The Public Corpus

- Selected parts of the reference corpus (approx. 50 hrs) will be made publicly accessible (including English translation and basic transcription/annotation)

Publication of the preliminary basic vocabulary

Publication of the public corpus

Publication of the first corpus based, electronic dictionary DGS–German



Poster presented at the Theoretical Issues in Sign Language Research (TISLR) 10 Conference, Sept 30 - Oct 2, 2010 at Purdue University, Indiana, USA.
The research leading to these results has received funding from the German Academies of Science Programme.

References

Hanke, Thomas / Hong, Sung-Eun / König, Susanne / Langer, Gabriele / Nishio, Rie / Rathmann, Christian: Designing Elicitation Stimuli and Tasks for the DGS Corpus Project. Poster presented at the Theoretical Issues in Sign Language Research (TISLR) 10 Conference, Sept 30 - Oct 2, 2010 at Purdue University, Indiana, USA.

Hanke, Thomas / König, Lutz / Wagner, Sven / Matthes, Silke: DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In: Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, LREC 2010, 22-23 May 2010, Malta. pp. 106-109.

König, Susanne / Konrad, Reiner / Langer, Gabriele / Nishio, Rie (2010): How Much Top-Down And Bottom-Up Do We Need To Build A Lemmatised Corpus? Poster presented at the Theoretical Issues in Sign Language Research (TISLR) 10 Conference, Sept 30 - Oct 2, 2010 at Purdue University, Indiana, USA.

Matthes, Silke / Hanke, Thomas / Storz, Jakob / Efthimiou, Eleni / Dimiou, Nassia / Panagiotis, Karioris / Braffort, Annelies / Choisier, Annick / Pelhate, Julia / Safar, Eva (2010): Elicitation Tasks and Materials designed for Dicta-Sign's Multi-lingual Corpus. In: Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, LREC 2010, 22-23 May 2010, Malta. pp. 158-163.

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: Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, LREC 2010, 22-23 May 2010, Malta. pp. 178-185.