DGS Corpus Project

Development of a Corpus Based Electronic Dictionary
German Sign Language / German

Siegmun Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, Arvid Schwarz

Public voting

- web-based voting system
- all interested members of the language community are entitled to vote (after registration)
- verification of signs of previously published sources and basic sign vocabulary
- validation of dialectal variants, sign forms and sign meanings
- approx. 350-400 hours
- 250-300 informants
- approx. 2.25 m. tokens

Focus group

- direct representatives of the language community
- approx. 20 deaf experts from different regions
- trained and sensitized to linguistic questions
- consulted in cases of doubt concerning, e.g. the lexical status of a sign, citation form, language change, sign use and (regional) variation
- decisions concerning lemma selection or the well-formedness of grammatical constructions continuously validated by focus group
- filling gaps, avoiding additional surveys

Analysis and compilation of dictionary entries

- abstracting from occurrences and all gathered data on a sign to obtain a description of lexical types
- description of regular grammatical properties of classes of lexical items
- description of forms (phonological variants, use of space, morphology, word class/syntactic functions, meanings)
- possible translations into German, iconic value & visualization technique, dialectal information
- cross references to related and similar signs, synonyms, antonyms, examples of use
- no documented empirical basis
- valuable source for basic vocabulary
- a pool of signs to be evaluated by linguists and native signers
- compilation of a preliminary basic vocabulary for DGS to be published after 5 years, ongoing validation by corpus data

State of the art:

- no large corpus of DGS available
- no comprehensive dictionary of DGS on linguistic principles available
- many sign collections of DGS available, differing in size, standard and degree of documentation
- several smaller project-based corpora from technical sign dictionaries available
- complex lexical database system (Lex) available
- experience and knowledge how from more than 10 years of dictionary making in Hamburg
- several other national corpus projects under way or in preparation
- cooperations, transfer/exchange of know-how, workshops

Data collection (Methods and Setting)

- at least 10 different locations (see map); areas of relatively high deaf population density (deaf schools, deaf centres...), which are easy to reach from surrounding rural areas, where mobile studio is successively set up in each of the locations
- pair of two informants
- peer-to-peer procedure
- interviews conducted by a local deaf contact person to secure an elicitation of regional sign variants with as little influence from the interviewer as possible
- each elicitation consists of:
  - (a) a standardized interview covering language and social data (approx. 20 min./informant); aim: description of data on metadata level → IMDI standard
  - (b) filming of spontaneous conversations on a given topic (2-3 topics from a list, 60-90 min. for each pair of informants); → elicitation of as much basic vocabulary as possible; list of topics covers about 20 subject areas.
- (iii) different tasks with selected stimuli (approx. 120 min. for each pair), Tasks aim at capturing special phonological, morphological, syntactic and lexical phenomena.
- setting: see illustration

Basic transcription

- translation into German
- segmentation into utterances
- contents accessible via German
- coding of mouting & mouth gesture, (lexical) facial expression, notation and classification of token form, contextual meaning, syntactic category, aspects of spatial use
- tagging of phrases and sign strings (within utterances)
- access via individual signs and special sign categories

Lemma selection and controlling of detailed transcription

- lemma selection on grounds of frequency of occurrence and distribution among signers, votes from focus group and public voting (see left side)
- approximately 6000 lemma entries expected
- preparation of information for focus group and public voting
- selection of passages for detailed transcription

Detailed transcription

- approx. 50% of the basic transcription
- differentiation of phonological variants, grammatical sign forms (e.g. plural, negation), modifications
- coding of mouting & mouth gesture, (lexical) facial expression, notation and classification of token form, contextual meaning, syntactic category, aspects of spatial use
- tagging of phrases and sign strings (within utterances)
- coding and annotation of productive sign forms
- more detail: transcription of surrounding context of a sign

Comprehensive, corpus based, electronic dictionary DGS-German

- contains approx. 6000 sign entries
- all contained information is corpus based
- entries contain descriptive information on signs
- entries include examples sentences taken from the corpus
- sign focus on DGS part
- German part provides access to DGS for hearing users: deaf users get basic information about German words
- combination of different types of dictionary functions
- primarily serves the following target groups:
  - DGS learners, e.g. hearing parents of deaf children, students of Deaf Studies/signe language interpreting
  - professional sign language interpreters for DGS - German, native signers of DGS: deaf adults, CODAs, deaf children or pupils, acquiring DGS as native language
  - sign language teachers, linguists and others concerned with sign language structure

Dictionary grammar

- based on the corpus and information from the focus group
- general overview of the most important grammatical features of DGS, supplemented by examples taken from the corpus
- written in easily understandable terms
- sign entries become shorter, more compact and therefore more clearly

Preliminary basic vocabulary

- basic vocabulary of DGS and German
- not based on corpus data but on published sign collections
- signs verified by focus group and public voting
- to be replaced by the general, comprehensive dictionary after 10 years

Dictionary including dictionary grammar

-Contains all contained information is corpus based
-Entries include examples sentences taken from the corpus
-Dictionary focus on DGS part

Analysis

- verification additional information on signs & uses
- analysis compiling entries
- dictionary including dictionary grammar
- publicly accessible corpus

Annotated public corpus

- representative parts of the corpus (approx. 50 hours)
- English translation of content, glosses and metadata
- suitable exchange formats provided (e.g. ELAN, IMDI)
- accessible online

Published sign collections

- no documented empirical basis
- valuable source for basic vocabulary
- a pool of signs to be evaluated by linguists and native signers
- compilation of a preliminary basic vocabulary for DGS to be published after 5 years, ongoing validation by corpus data


data collection
basic transcription
lemma selection, analysis & compilation of dictionary entries
public voting & consultation of focus group

Poster presented at the LREC 2008 Conference in Marrakech, June 1st, 2008