DGS Corpus Project – Development of a Corpus Based Electronic Dictionary German Sign Language / German

Siegmund Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, Arvid Schwarz
Institute of German Sign Language and Communication of the Deaf (IDGS), University of Hamburg
Binderstr. 34, 20146 Hamburg, Germany
{Siegmund.Prillwitz,Thomas.Hanke,Susanne.Koenig,Reiner.Konrad,
Gabriele.Langer,Arvid.Schwarz}@sign-lang.uni-hamburg.de

Abstract

This paper introduces a 15-year project, which aims to combine the collection of a large corpus with the development and production of a comprehensive, corpus-based electronic dictionary of German Sign Language (DGS).
The scope, aims and the methodological approach of this large-scale project, accepted for funding by the Hamburg Academy of Sciences, are discussed.

1. Introduction

In modern lexicography there is an increasing awareness that dictionaries should be corpus-based. In sign language lexicography only few corpora exist and none of the larger corpora have been thoroughly transcribed and analysed for dictionary production. This gap is going to be filled by the project presented here. Different sources are used first to collect language data and later to analyse and validate the data on an empirical basis. The primary source is a sign language corpus to be built during the project. A voting system as well as a focus group complement the corpus data. In addition, previously published sign collections are taken into account.
The corpus is completely annotated and lemmatised. Further analysis and detailed annotations of selected lemmas are made particularly with regard to the compilation of the dictionary and the dictionary grammar. If necessary, others of the sources mentioned above support the ongoing analysis.
The primary products are a large reference corpus, partly published with English translation, and a comprehensive, corpus-based electronic dictionary of German Sign Language (DGS) – German, including a dictionary grammar. The dictionary will be preceded by a preliminary collection of basic vocabulary of DGS published after five years.
In the following, we outline properties of the corpus and the elicitation settings as well as the different annotation processes. The various sources of information we use, and their respective functions are introduced and distinguished. Features of the analysis up to the composition of the dictionary entries are presented. Finally, the products resulting from the project are briefly described.

2. Corpus design

A corpus of approximately 350–400 hours from 250–300 informants will be collected. We anticipate a number of approx. 2.25 m. tokens. This is, in size and scope, comparable to large spoken language corpora.
The corpus should reflect a representative and well-balanced part of everyday communication of competent deaf signers. Informants from all over Germany of different age, sex and social status are included. They are selected on grounds of their language competence, their membership in the deaf community, and their regional rootedness. For logistic reasons the elicitations are carried out in ten cities, all of which are areas of relatively high deaf population density and are easy to reach from surrounding rural areas. A mobile studio is set up and is moved from place to place during the first three years of the project.

Figure 1: Preliminary set of locations

The design allows the use of the corpus for various tasks, such as the validation of the basic vocabulary, thorough research on DGS grammar based on the transcriptions and the identification of different meanings and collocations of a sign by appropriate contexts. Furthermore, the design anticipates a comparative sociolinguistic study compara-

ble in kind and quality to Lucas et al. (2001) and Schembri/Johnston (2004).

3. Elicitation

The elicitations follow best current practice, considering not only academic but also social, political, ethical and legal aspects. To describe filmed material on a metadata level, we make use of the IMDI standard (IMDI 2003, Crasborn/Hanke 2003)

Pairs of two informants are appointed for each elicitation session. The elicitations are carried out using a peer-to-peer procedure where two informants change roles according to situations. The interviews are conducted by a deaf contact person from the respective region to secure an elicitation of regional sign variants with as little influence from the interviewer as possible.

The elicitation consists of

- a standardized interview covering language and social data (approx. 20 min./informant)
- the filming of spontaneous conversations on a given topic (selected from a list of topics, 60-90 min. for each pair of informants)
- different tasks with selected stimuli (approx. 120 min. for each pair).

The major aim of the spontaneous conversation is to elicit as much basic vocabulary as possible, i.e. signs, which are used in every-day situations on a regular basis. Therefore, a list of topics is supplied, covering about 20 subject areas. From this list, the informants are asked to select 2-3 topics.

The different tasks with selected stimuli aim at capturing special phonological, morphological, syntactic and lexical phenomena. This explains why the stimuli tasks are given more time compared to the spontaneous conversations. We use stimuli from already carried out or planned elicitations and, if necessary, adapt them for DGS in order to enable parallel cross-linguistic analysis.

The informants are seated more or less facing each other, following a preliminary survey of informants' preferences. Each informant is filmed by a high-definition camera, two additional DV-cameras allow for a bird's eye view for an in-depth analysis of the use of space and a third camera films the whole scene to get a complete picture of the informants' interaction with each other.

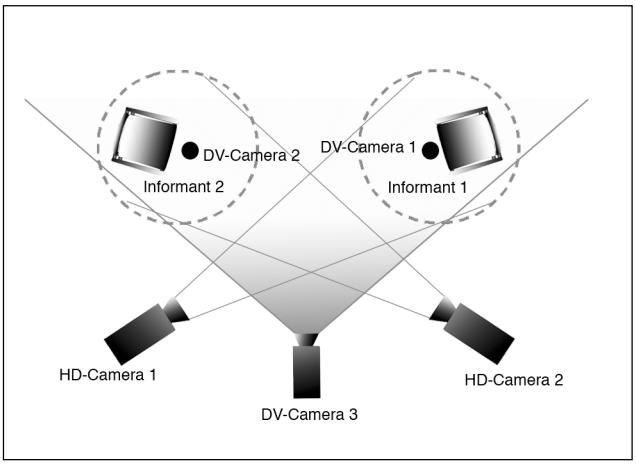

Figure 2: Elicitation setting

Elicitation sessions are expected to last approximately four hours. We anticipate about 80 min utilisable video material per informant.

The elicitation process is recorded, the obtained data is digitised on-site and specific (esp. linguistic) features are noted. The films are successively made available for the basic transcription.

4. Annotation

The corpus data will undergo different annotation and transcription processes aimed at identifying signs and documenting their properties. After a translation of the conversations into written German, a basic transcription serves to segment utterances and to identify lexical items and thus to provide a first access to the data. Second, approximately 50% of the transcriptions will be transcribed again in more detail.

4.1 Translation and Segmentation

As a very first step the elicited DGS material is translated into German by trained interpreters and synchronised with the DGS texts. Thus, the content of the interviews is captured and becomes searchable via written German. The alignment also provides a first rough segmentation of the signed text into meaningful sections. Furthermore passages which are special with respect to technical, linguistic or content matters are marked and documented in a report by the interpreters. These remarks, among other criteria, are taken into account when assigning priorities for further analysis to certain sections. They are also taken into consideration when selecting parts of the corpus for the detailed transcription and for publication.

4.2 Basic transcription

The basic transcription serves to segment utterances and to identify lexical items. It provides a first and easy access to the individual signs. The previous segmentation resulting from the translation is reviewed and refined from a sign language perspective. The transcription is carried out according to guidelines and criteria described in a manual. Tokens are assigned to types which are collected in a lexical database that is part of the iLex transcription environment (cf. Hanke/Storz, this volume). Each type is labelled by a unique gloss and its form is described by means of a HamNoSys notation. During the transcription, interesting or special passages are marked for further analysis.

At this stage of transcription, productive sign tokens as well as special signs (such as index, manual alphabet, etc.) are assigned to larger groups and coded only in a very broad way. For productive signs, such groups are determined by handshape/hand configuration and visualisation techniques (cf. Langer 2005).

4.2.1 Lemma selection and management of detailed transcription

After the basic transcription, the data are accessible from a type-token perspective. This allows lemmas to be selected for the dictionary and stretches of text as well as single tokens to be assigned to further detailed transcription.

One major criterion for the inclusion of a sign into the dictionary is the frequency of its occurrence and its distribution among the pool of informants. Other criteria considered include votes from the focus group or the deaf community (see below: public voting). Taking into account the structural characteristics of signed languages and the current state of sign language lexicography, 6000 entries look like a realistic number for a comprehensive DGS sign language dictionary (cf. König, Konrad, Langer, in preparation).

For each selected sign, a linguist will review all occurrences and decide on whether all of them or just selected tokens have to be transcribed in more detail. For this decision the number of tokens, the preliminary assessment of the sign kind (i.e. invariant or highly modifiable) and the probable differentiation of its meanings are taken into consideration. Other reasons influencing the selection for detailed transcription are the suitability of utterances as dictionary examples, and instances of grammatical phenomena, which can serve as a basis for the compilation of the dictionary grammar. A detailed transcription allows for a closer look and well-founded analysis of these grammatical phenomena.

4.3 Detailed transcription

In the second transcription stage, selected occurrences of sign lemmas are transcribed again in more detail and with their surrounding context. We expect that approximately 50% of the basic transcriptions will be refined in more detail. The annotation and transcription of the corpus will be closely intertwined with the requirements of the lexical analysis needed for dictionary production. Not only a given occurrence, but also the surrounding utterance of a token is transcribed in order to be able to pin down the contextual meaning of the sign, collocations and other relevant context information. The following aspects are going to be coded:

- mouthing or mouth gesture,
- (lexical) facial expression,
- notation and classification of the form of the token,
- contextual meaning of the token,
- syntactic category,
- aspects of spatial use: establishing of spatial scenes, positioning of objects at special places in signing space and relating back to established places,
- productive elements for the visualisation of objects, processes, etc.

During the detailed transcription, utterances are segmented into smaller units (phrases, sign strings). Thus, sentence analysis regarding functions of single signs (e.g. illustrative function) is made possible. Interesting passages with regard to content or language are, again, noted in a report. Short passages that can be well understood without further context are marked as potential example sentences or references for the dictionary.

Types are also further differentiated, their tokens are described more closely and classified, e.g. into

- phonological variants,
- forms, which are the result of grammatical processes, such as spatial orientation or positioning, marking of aspect, plural marking and others,
- forms, which result from sign modification, e.g. as a result of metaphoric use or of re-iconisation.

Productive signs are an important component of signed texts and have to be taken into consideration during the process of detailed transcription. Alongside the established lexicon, these have to be coded and annotated appropriately.

4.4 Transcription team

By opening the field of professional transcription work to students at the IDGS on such a large scale, we enter new ground. Both, basic and detailed transcriptions are mostly carried out by students. The students are trained and constantly supervised by experienced deaf transcribers. This procedure contains several advantages. In such a process, the transcriptions are continuously checked and verified by a team of experienced native signers, i.e. transcriptions are looked at by at least two persons working independently. This ensures a high quality of transcription without doubling the costs. For the students, we provide a first access to corpus linguistics and a first-hand insight into practical sign language lexicography, which is resumed by a long-term perspective via a PhD position. In addition, the combination of experienced deaf transcribers and students allows us to transcribe larger amounts of data within the timeframe of the project than by staff transcribers only.

iLex models the whole the transcription process, and especially supports the consistency of token-type matching. Transcribers are enabled to communicate with each other within the transcription environment by means of video chat and web 2.0 technologies, contributing to a steady flow of information within a rather large group of transcribers.

5. Data Sources

As outlined at the beginning of this paper, we rely on different sources to gather, verify and analyse language data, all of which complement each other to different degrees. Additionally, language knowledge and intuition of the deaf team members come in at all points. In the following, the four sources and their main functions are briefly introduced.

5.1 Published sign collections

For DGS there are many published collections of signs, differing in size, standard and degree of documentation (e.g. Metzger et al. 2000, 2003; *777 Gebärden* 2002;

Grundgebärden 1 & 2 1999, 2000). Nevertheless, these are a valuable source for basic DGS vocabulary as it can be assumed that many of the published one-to-one translations of German basic vocabulary consist of agreed-upon signs, which are actually in use in the language community. The function of such sign collections in the project is to provide a pool of signs, which are evaluated by linguists and native signers to compile a preliminary basic vocabulary for DGS on this base.

5.2 Corpus of natural language

The corpus, which is elicited during the first phase of the project, serves as the major source to draw information from and is the very heart of the project. Lexemes can be identified and annotated with regard to the dictionary and the empirically based dictionary grammar. Every corpus, however, is limited. Not every topic can be covered in the elicitations, especially spontaneous conversation is unpredictable. Lexical gaps in the corpus, especially for everyday contents, can be searched for in the sign collections presumably with a high rate of success. Also, not all kinds of grammatical and lexical phenomena can be assumed to be contained in the corpus if not particularly asked for. Data from published sign collections, can be verified or discarded on grounds of corpus data to compare with.

5.3 Focus group

Additional elicitations later on in the project are costly and time consuming. For cases of doubt concerning, for example, the lexical status of a sign, matters of language change, different meanings or regional information, a focus group is installed. The focus group consists of approximately 20 deaf experts from different regions, which are trained and sensitised to linguistic questions. The members of the focus group are direct representatives of the language community and their active involvement is vital for the success of the project. Decisions such as those concerning lemma selection or the well-formedness of grammatical constructions will be continuously validated by the focus group. By falling back on the focus group, gaps in the corpus can be compensated for and additional surveys can generally be avoided. In some cases, it may still prove useful to conduct small additional surveys or elicitations to clarify certain questions or to supplement the data. Members of the focus group can then act as contact persons in their respective regions.

5.4 Public voting

A general voting web interface is established, which is open for all interested members of the deaf community. The focus group is very limited in terms of the number of members and, expectably, cannot answer all questions. The feedback from the focus group or partial results are still valuable as a starting point for public voting. Furthermore, the voting is useful for the validation of dialectal variants; variants that have not been included so far can be put in for discussion or for further voting by deaf users, respectively. Signs which are only contained in the sign collections used for the basic vocabulary but not in the corpus can be validated by means of public voting as well.

Public voting allows a substantially larger part of the language community to directly participate in the project than possible through the informants, the focus group, and team members.

6. Analysis and composition of dictionary entries

The dictionary will be entirely based on the corpus with respect to the signs to be included as lemmas. Most examples of sign uses will be taken directly from the corpus. However, the information provided in the entries will decidedly exceed a conglomeration of just corpus references and examples. Rather, we will systematically abstract from the occurrences to obtain a generalized description of lexical items. This description will include following aspects:

- sign form (citation form)
- phonological variants
- use of space, morphology of sign (e.g. plural forms), modifiability of the sign
- word class, syntactic functions
- meaning: different readings; possible translations to German
- iconic value and visualisation technique; popular explanations as known in the sign community
- dialect information
- cross-references to related and similar sign, synonyms and antonyms
- examples to illustrate grammar, usage and different readings of the meaning: examples taken from the corpus or, where needed, invented examples, verified by the focus group.

After a first preliminary analysis, it is determined which additional information is needed for the final analysis and the composition of the entry. Additional detailed transcriptions are requested and material is given to the focus group and prepared for public voting. Once the detailed transcription of the requested sections is completed, large quantities of data can be compared and processed for the dictionary entry.

Not only is the structure of the lexicon analysed and described, but also grammatical phenomena are looked at more closely in order to compile a corpus-based dictionary grammar. To this aim, the filmed material from the tasks with different stimuli is used predominantly, but also grammatical constructions in fluent discourse (i.e. from the conversations) are examined.

7. Products

Several products will result from this project: a preliminary collection of basic DGS vocabulary, a large research corpus for DGS, a publicly accessible part of the corpus (supplied with English translation), the dictionary

grammar and a number of academic publications accompanying the work in progress.

7.1 Corpus-based electronic dictionary

The dictionary will be the first comprehensive corpus-based dictionary of German Sign Language. It is published in electronic form and primarily serves the following target groups:

- DGS learners who are native speakers of German, e.g. hearing people dealing with deaf people in work related contexts, parents and relatives of deaf children, students of Deaf Studies/sign language interpreting, and hearing impaired or late deafened adult learners of DGS as a second language,
- professional sign language interpreters for DGS ↔ German,
- native signers of DGS: deaf adults, children of deaf adults (CODAs),
- deaf children or pupils, acquiring DGS as native language,
- sign language teachers, linguists, language typologists and others concerned with sign language structure on a theoretical or practical level.

In order to be able to serve such diverse groups, the dictionary has to combine different types of dictionary functions to allow various kinds of uses. Lemma selection and lemma articles clearly focus on the DGS part. Like in a learner's dictionary, example sentences are included to illustrate sign uses and meanings. The German part first of all provides access to DGS for hearing users via their native language. Additionally, deaf users get basic information about German words and example sentences. It thus serves as a first starting point for the consultation of further references. The dictionary is going to contain an estimated number of 6000 sign entries.

7.2 Dictionary grammar

An important part of a comprehensive dictionary is the dictionary grammar. Regular grammatical properties of lexical items and formal paradigms do not have to be listed time and again in the entries. References to the dictionary grammar render the entry shorter, more compact and therefore make it more clearly. Learners also greatly benefit from a solid and well-founded grammar of the target language written in easily understandable terms. The dictionary grammar will be based on the corpus; issues, which cannot be resolved by solely looking at the corpus data, are given to the focus group or small additional elicitations have to be carried out. It will contain a general overview of the most important grammatical features of DGS, supplemented by examples taken from the corpus.

7.3 Annotated corpus

To comply with international corpus-linguistic objectives, a representative selection of about 50 hours of corpus material will be made publicly available online. To make the data accessible for researchers who do not understand German, an English version of the existing German translation is provided for this part of the corpus. All annotated signs are additionally endued with an English gloss. iLex provides formats for the interchange of transcripts and metadata without loss.

In the context of co-operations, the complete corpus will be available for linguists and PhD candidates, working on special issues in exchange for supplying additional annotations which can, in turn, be used by the project team.

7.4 Preliminary collection of basic vocabulary

In the fifth year of the project, a collection of basic DGS vocabulary will be published in an electronic version. This dictionary is preliminary since it is not based on corpus data. It is planned as a bilingual dictionary for DGS and German and should include the basic vocabulary of both languages. To compile this dictionary, already published sign collections will be compared and evaluated. Signs that are listed several times in different sources are likely to be included in the basic vocabulary, signs that have only one or few listings have to run through a verification process (by the focus group and the public voting) to avoid a listing of artificial signs and artefacts. A basic vocabulary list for German will be included, and missing sign equivalents will be provided by different methods such as small elicitations, public voting and input from the focus group. At the end of the project this product is replaced by the larger, fully corpus-based dictionary.

8. Co-operation

Besides elicitation settings, tasks and technical equipment mentioned above, we consider it essential to push standardisation or at least compatibility of annotation and transcription conventions to reach comparability of results across projects in cross-linguistic research. To this aim, we have arranged co-operations with other national corpus projects and look forward to co-operating with more projects currently in preparation. Topics like quality assurance or transcription efficiency will be best discussed in cross-project workshops.

9. References

777 Gebärden 1-3. Alle 3 Folgen auf einer DVD (2002). Version 2.0. Guxhagen: Manual Audio Devices. (DVD-ROM).

Crasborn, O., Hanke, T. (2003b). *Additions to the IMDI metadata set for sign language corpora*. Agreements at an ECHO workshop, May 8–9, 2003, Nijmegen University. URL: http://www.let.ru.nl/sign-lang/echo/docs/SignMetadata_Oct2003.pdf (last accessed March 2008).

Grundgebärden 1. Für Einsteiger. (1999). Hamburg: Verl. hörgeschädigte kinder. (CD-ROM).

Grundgebärden 2. (2000). Hamburg: Verl. hörgeschädigte kinder. (CD-ROM).

IMDI (ISLE Metadata Initiative; 2003). *IMDI Metadata Elements for Session Descriptions, Version 3.0.4*, MPI Nijmegen. URL: http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.4.pdf (last accessed March 2008).

König, S. Konrad, R, Langer, G. (in preparation). Lexikon: Der Wortschatz der DGS. In H. Eichmann, J. Hessmann. *Einführung in die Gebärdensprachlinguistik*. Hamburg: Signum.

Langer, G. (2005). Bilderzeugungstechniken in der Deutschen Gebärdensprache. In *Das Zeichen* 70, pp. 254--270.

Lucas, C., Bayley, R., Valli, C. (2001). *Sociolinguistic Variation in American Sign Language.* Washington, DC: Gallaudet Univ. Press.

Metzger, C., Schulmeister, R., Zienert, H. (2000). *Die Firma. Deutsche Gebärdensprache do it yourself.* Hamburg: Signum. (CD-ROM).

Metzger, C., Schulmeister, R., Zienert, H. (2003). *Die Firma 2. Deutsche Gebärdensprache interaktiv. Aufbaukurs in Deutscher Gebärdensprache – Schwerpunkt Raumnutzung*. Hamburg: Signum. (CD-ROM).

Schembri, A., Johnston, T. (2004). Sociolinguistic variation in Auslan (Australian Sign Language). A research project in progress. In *Deaf Worlds* 20 (1), 78--90.