Corpus-based Sign Dictionaries of Technical Terms – Dictionary Projects at the IDGS in Hamburg

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

Abstract

In this paper we give an overview on the six corpus-based sign language dictionaries of technical terms produced by the lexicographical team at the IDGS in Hamburg. We shortly introduce the different work steps. Then we focus on those work steps, which deal with or rely on corpus data. The consistent token-type matching and annotating accomplished during the transcription process allows for comparing the transcribed answers and for evaluating them quantitatively and qualitatively. Based on this analysis appropriate signed translations of technical terms are selected. In the dictionaries all single signs included in the selected answers are listed and described as they would be in a general sign language dictionary. During the process of transcription, selection and analysis assumptions and practical decisions have to be made. We discuss some of the assumptions and decisions that have proven valuable over time, as well as some open questions.

1. Projects

At the Institute of German Sign Language (IDGS), six dictionary projects in such diverse technical fields as computer technology, psychology, joinery, home economics, social work, and health and nursing care have been carried out. A seventh project on landscaping and horticulture is in progress.

Six of the seven dictionaries are based on corpus data collected from deaf experts in the respective fields. Elicitation methods, such as interviews and picture prompts, corpus design as well as annotation, transcription, sign analysis and dictionary production have been continually developed and refined over the years. Many procedures rely heavily on the use of a relational database (iLex; see Hanke & Storz, this volume).

The following table provides an overview on the the six projects and their elicited corpus data:

	Psychology	**Joinery**	**Home Economics**	**Social Work**	**Health and Nursing Care**	**Landscaping and Horticulture**
Timeframe	1993-1995	1996-1998	1998-2000	2001-2003	2004-2007	2006-2009
Number of technical terms	900	800	700	450	1000	710
Number of signed translations included in the dictionary	1270	2800	1560	940	2330	
Stimuli:						
• written terms	900	800	700	450	1000	710
• pictures	0	550	340	0	190	410
informants (filmed)	5	16	17	15	18	11
informants (transcribed)	5	10	11	10	10	
Hours of filmed material						
• interview	2	3,5	2	5	5	3,5
• conversation	12	19	15	9	8,5	5,5
• elicitation	7	32,5	37,5	40,5	93,5	37
answers (total)	3600	13500	12500	9600	43200	21100
answers (transcribed)		8900	9800	6800	15200	
Number of transcribed tokens (single signs)		18700	26350	15800	29500	
Number of						
• types		1370	1750	1766	1450	
• productive signs		2800	2850	50	2300	

Table 1: Figures of technical sign dictionaries and corpus data.

Each project is completed within a timeframe of about 2,5 years which allows for a coverage of 500 to 1000 technical terms. In order to provide DGS equivalents to technical terms a corpus-based and descriptive approach has been chosen. Nearly all technical content has been produced in cooperation with experts from educational or

academic institutions of the respective field. These experts compile a list of technical terms, write the definitions for these terms and produce appropriate illustrations. All lexicographic work concerning eliciting, transcribing, analysing, presenting signed translations and single signs, and producing the actual dictionary is carried out by the lexicographical team at the IDGS. From 1996, the core team has consisted of four to six deaf and three hearing colleagues. Most team members have been working in the dictionary projects since 1996. This has facilitated a continuity of experience and know-how as well as a continuous improvement of methods and procedures.

2. Work steps

The following table outlines the main work steps in our empirical approach, following a roughly chronological order. All tasks concerning technical and terminological information (e.g. definition, illustration, subject categories, synonyms) which are executed by experts in the field working in vocational or academic institutions are not listed in table 2. Also, the production steps are left out.

Work step	**Tasks and procedures**	**Progression and results**
(1) Data collection		
(1a) Preparation	• Searching for deaf informants (fluent DGS signers trained and working in the field) • Testing equipment and studio setting • Elicitation material	Word list (ordered by subject categories) with context information, combined with illustrations
(1b) Data collection	• Interview (standardised) • Interview (pre-structured) • Elicitation (written terms and pictures as stimuli)	• social and linguistic background (meta-data) • conversational data • spontaneous responses
(2) Definition of the corpus		
(2a) Documentation and segmentation	• Formatting digitised material • Linking films to database (iLex) • Conversational data: Segmentation and description of content; Tagging (linking to terms) • Elicited data: Segmenting in subject catogories; Tagging (linking to terms)	QuickTime® movies Content: search by written German; direct access to DGS equivalents via terms Direct access to all answers via terms
(2b) Review of data	Conversational data: • Qualitative evaluation of informants' DGS competence Elicited data: • Tagging repetitions, wrong, and odd answers for exclusion • Annotating informants' and transcribers' judgement of the answer	Priority list for transcription Defined corpus for transcription (including conversational data) Documentation of appropriateness of answers for selection process
(3) Transcription and annotation		
(3a) Token-type matching	• Identification of lexemes, variants and modifications • Identification of productive signs and others (e.g. numbers, indexes, manual alphabet etc.)	Direct access to tokens via types and vice versa
(3b) Annotation	• Form (HamNoSys) for types (citation form) and tokens (variation, realisation in context, deviation) • Mouthing • Meaning (types and tokens)	Search by HamNoSys Search by mouthing Search by meaning
(4) Selection of signed translations		
(4a) Selection	Selection of answers and DGS equivalents as translations of the corresponding technical term	DGS translations of technical terms
(4b) Filling gaps	New combinations of signs and mouthings or coining new signs	

Work step	Tasks and procedures	Progression and results
(5) Analysis of conventional and productive signs used in the selected DGS translations		
(5a) Conventional signs	• Empirical status • Sign form (citation form, variants, modifications) • Meaning • Iconic value and visualisation technique • Use of signing space • Similar and related signs (synonymous, homonymous signs) • Comments (e.g. dialect, variation of form)	Lexical analysis and description of lexemes and productive signs
(5b) Productive signs	• Iconic value and visualisation technique • Similar and related signs	
(5c) Quality control	• Consistent token-type matching • Constistent description of types and productive signs	

Table 2: Work steps focussing corpus-related tasks.

3. Corpus-related tasks

Data collection, reviewing and annotating are time-consuming procedures. Due to the timeframe of 2-3 years for each dictionary project, annotation and transcription is restricted to the elicitations and conversational data of about 10 informants. This means that the corpus represents a relatively small section of all existing or possible translations of technical terms in DGS discourse. Nevertheless there are some striking arguments in favour of a corpus-based approach:

- The selection process can be based on the frequency of elicited answers.
- The transcribed data show the variety of signs, sign-mouthing and sign-sign combinations. This provides a solid basis for assumptions on sign formation and sign structure, and for decisions in the lexicographic process.
- All decisions can be traced back to the original data which allows for revision of transcription and lexical analysis.

From the joinery project on the corpus data contained suitable translations for almost every technical term so that newly coined signs make up for less than 1,5% of all given translations in each dictionary.

Over the years, elicitation techniques, documentation, segmentation, annotation and transcription have been developed and refined. Corpus-related tasks are the definition of the corpus (reviewing of data, see table 2: step 2b) and transcription (token-type matching and annotation, see table 2: step 3). Also, the selection of elicited answers (see table 2: step 4) as well as the lexical analysis and description of conventional and productive signs (see table 2: step 5) require transcribed and annotated data and are thus strictly corpus-related. In the following, we describe the tasks and procedures of these work steps in more detail.

3.1 Review of data

The pre-structured interviews are segmented in question (interviewer) and response tags (interviewee) and the contents are translated or summarised in written German. Further, sequences that correspond to technical terms are tagged so that spontaneous conversational DGS equivalents and elicited answers can be transcribed and easily compared to each other (see table 2: step 2a).

The signers' DGS competence and signing styles are evaluated on ground of the conversational data. Deaf colleagues check the following aspects of signing: general impression (naturalness and fluency of signing, comprehension), context and text structure, grammar, lexicon, mouthing, facial expression, reference to technical terms. The evaluation results in a priority list determining which informants will be transcribed first.

All elicited answers of each informant are linked to the corresponding technical term and reviewed by a second deaf colleague. First, the appropriateness is assigned to the given response. Valid answers are selected for transcription. Wrong or odd answers with regard to content and form (e.g. slips of the hand) are excluded from transcription. Also repeated, identical answers are marked and excluded from transcription. Second, the informants' judgements of their answers are documented and the answers are evaluated by the transcriber. For example: the informant shows that he is doubtful or feels incomfortable with his signing, he wants to correct the answer, or he does not have a valid translation but wants to make a proposal. Even if the answer is spontaneous, the transcriber can judge the response as due to the elicitation setting and not likely to occur in natural signing, as a proposal or as an atypical DGS construction. Annotating informants' and transcribers' judgements provides important clues for evaluating the tagged and transcribed answers during the selection process.

3.2 Transcription and annotation

During the transcription process, conventional signs (lexemes), their variants and modifications, as well as productive signs and other instances of signing, such as indexes, numbers or fingerspelling are identified, classified and annotated.

Tokens are compared to each other and to already existing types. Similar tokens with regard to form and

meaning are grouped together and matched to types (token-type matching). Types are differentiated from each other with regard to form, iconic value, visualisation technique and meaning. The citation form of a conventional sign is determined on the basis of the matched tokens and described via HamNoSys. Deviations of token forms from the citation form are also documented. Variants and modifications are treated as separate but related types labelled by the same gloss with different additional specifications (cf. König, Konrad & Langer, in preparation). The mouthing accompanying each sign or sign string is documented. Mouthings help to determine the signs' meanings. In most, but not in all cases, the meaning corresponds to the technical term in question (cf. Langer, Bentele & Konrad 2002).

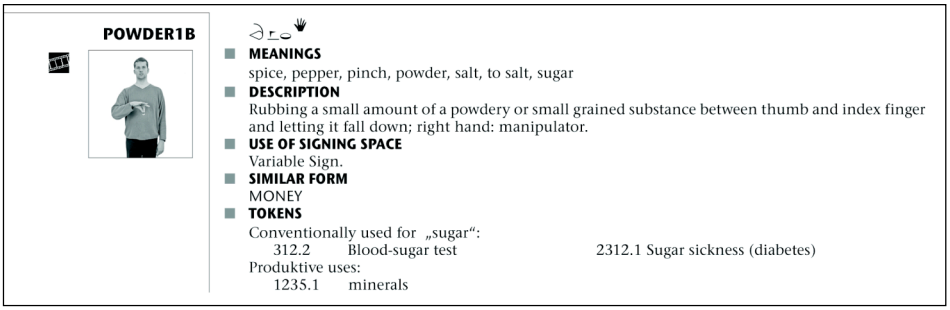

Figure 1: Sample entry of a conventional sign

3.3 Selection of signed translations

All signed translations of technical terms included in the dictionary are taken directly from the corpus. iLex provides a comprehensive view of all answers to one term, also showing gloss strings and further annotations. Identical responses of different signers are easily detected by sorting the answers by gloss string. The database allows for a very quick and direct access to the original data so that for the selection the original film sequences can be viewed to verify the annotations. Frequency of occurrence and wide distribution among different informants is an important criterion for selection. The selected answers consist of:

- conventional signs and sign combinations of conventional signs, including modifications and productive sign-mouthing combinations),
- productive signs transporting the meaning in a clear and striking image,
- a combination thereof.

Several acceptable answers may be selected to display different variants of signs or sign combinations found in the corpus.

If no acceptable translation is found, a new sign or sign combination is created. For filling these gaps the deaf colleagues ask one of the deaf informants to make a proposal or to discuss their own ideas. In many cases new sign combinations include single conventional signs or productive signs taken from the corpus. New sign combinations or newly coined signs are labelled as such in the dictionary. Except for the psychology dictionary which was the first corpus-based project with a very small data collection, sign creation is marginal compared to other tasks (see above).

3.4 Analysis of lexemes and productive signs

From 1998 on, the dictionaries include an inventory of single signs used in the translations of the technical terms. Each sign is listed in a separate entry, ordered by glosses. The structure of these entries is similar to what you would expect from a general sign language dictionary. For each conventional sign, form (picture or film and HamNoSys), meaning, iconic value (image description) and visualisation technique, use of signing space and cross-references to similar or related signs are given. Occurrences of the sign in the DGS translations are listed at the end of each entry under the heading "TOKENS".

iLex allows for a quick access to all tokens grouped together in one type. A type in the database corresponds to a conventional sign, a productive sign or other sign categories such as numbers indexes, or fingerspelling. For ease of handling modifications and stable variants of conventional signs, as well as very similar instances of productive signs, are also grouped as types. Modifications are defined as a change of form as result of exploiting the iconic value of a sign in order to express a more specific meaning. Occurrences found in the project corpus and in transcriptions of other projects using iLex are taken into account for the lexical analysis. However, due to the limited size of the corpus and reduced context information of elicited answers, not all information given in the sign entries of the dictionaries is validated by the corpus. Other sources such as deaf colleagues' knowledge and intuition and small informal surveys have been used to supplement lacking corpus data.

3.4.1 Empirical status

We differentiate between productive and conventional signs. Criteria for the identification of conventional signs are frequency of use, distribution among signers, conventionalised and thus stable form-meaning combination, conventionalised association with a mouthed word. The latter are considered conventional uses of a sign, in contrast to productive uses where the same sign is combined with an occasional mouthing to express a specific meaning. The frequency of use of conventional signs across the informants is documented by a symbol on the left side of each sign entry.

3.4.2 Sign form

In many cases the corpus provides several identical realisations of conventional signs from which a citation form can be drawn. Also co-occurring stable variant

forms can be identified. Instances of sign modification, orientation and location in the signing space are related to the basic form by means of glossing conventions or by annotating the deviation in form of the token using HamNoSys.

3.4.3 Meaning

Conventionalised meanings of signs are frequently used and widespread across signers. Even out of context, the sign's form is associated with a certain meaning. Many conventional signs are combined with a mouthed word that corresponds to the intended meaning. Due to the elicitation method of using written stimuli, mouthings may occur more often than in natural DGS discourse. The informant may also be tempted to produce spontaneous sign-mouthing combinations which we consider as productive uses of conventional signs. As a third effect many responses to German compounds are sign strings that follow the sequence of the compound parts. Referring to words by mouthing and by combining signs to sign strings, are common strategies in DGS to express specific meanings, especially those of technical terms. The problem, however, is to determine the well-formedness and the degree of conventionalisation of these constructions. As long as there is no reference corpus of natural DGS discourse, decisions are primarily based on native signers' intuition.

Many signs are polysemous, i.e. one sign is used to express different meanings. This phenomenon is reinforced by the combination with different mouthings. In general, these meanings can all be related to the underlying image of an iconic sign.

In addition, the interplay of mouthing and iconicity is one reason for a high degree of lexical variation (synonymy). In different sign forms different aspects of the extra-lingustic referent can be visualised. In DGS there are, for example, at least four conventional signs for the meaning 'garden'. Two signs visualise raking (one with bent fingers representing the rake by substitutive technique, the other representing the hands handling a rake by manipulative technique), another digging (flat hand representing the blade of a spade; substitutive technique) and a fourth sticking seeds or cuttings into the ground (hands representing the hands holding small objects; manipulative technique).

3.4.4 Iconic value and visualisation technique

The iconic value of a sign cannot be directly and exclusively determined on the basis of corpus data. However, some evidence can be drawn from the corpus. Especially modifications, which exploit the iconic value of the basic sign, can be helpful. Possible modifications can also provide clues to the visualisation technique employed. Formational elements of a sign need to be set in relation to its conventional meanings and to be compared to productive sign use, existing variants of the sign and related signs with similar forms and underlying images. For these checks corpus data serve as a reference. In addition, signers' popular explanations may also provide valuable hints to describe the underlying image of a sign or its parts.

The iconic value is a valid criterion to distinguish sign homonyms. Signs with the same form and different underlying images are considered to be homonyms.

Iconic signs can be classified according to the visualisation technique involved (Langer 2005, König, Konrad & Langer, in preparation). Most of the signs can be analysed by three different techniques:

- Substitutive technique: The hand stands for the object or a significant part of the object e.g. the flat hand, palm down, represents a vehicle moving along a path.
- Manipulative technique: The hand stands for the hand of a person touching, holding or using an object, e.g. the fist represents the hand holding something.
- Sketching technique: The hand or part of the hand – e.g. the fingertip – works like a drawing tool, just like a pencil or a brush, tracing the shape of an object into three-dimensional space, e.g. the index finger is used to draw a circle.

3.4.5 Use of space

In order to enable the user to use a given sign in context, we provide information on how the sign can be modified by exploiting the signing space. Signs are divided into four categories:

- Invariable signs: The form of such a sign is fixed and cannot vary without becoming incomprehensible or changing the meaning of the sign. Most of these signs are body-anchored.
- Variable signs: The forms of most signs can vary by being orientated or located in space.
- Variable body location: This sub-group of variable signs can be modified by changing their body locations to express a more specific meaning. For example, in the citation form of the DGS sign for 'blood' or 'to bleed' the dominant hand starts near the palm of the non-dominant hand. A change of location, starting at the shoulder, means 'blood or bleeding at the shoulder'.
- Variable body and space location: Some signs can change location in space or on the body to express different meanings, e.g. with the open 5-hand one can mark a specific area in space or on the body.

3.4.6 Similar and related signs

Cross-references to homonymous signs and signs with similar forms are given in the sign entries. These references help to understand how forms, meanings and underlying images are interconnected and to become aware of similar signs that are not to be confused with the given sign. For example, as mentioned above some signs can be analysed as a lexicalised modification of a basic form.

In iLex cross-references to types of same and similar form can be used as another way of access to sign forms when searching a certain type by sign form during the

transcription process.

3.4.7 Comments

For some signs, additional information is given concerning specific use or aspects of the sign form. Dialectal variants can be identified by analysing the distribution of sign uses with regard to affiliations of informants to dialectal regions. Due to the relatively small sample sizes of our corpora, definitive dialectal surveys cannot be conducted. However, in some cases there is good evidence in the corpus for marking these signs as regional dialects.

Further, in some cases it is hard to decide whether the citation form is one- or two-handed, with or without repetition or if a circular movement is executed clockwise or anti-clockwise in the standard form. If corpus data suggests that these forms co-exist in free variation, a comment is added to the sign entry.

4. Assumptions and practical decisions

The central problems of analysing signed equivalents of technical terms are the identification of sign lexemes and the token-type matching, which require practical decisions based on theoretical assumptions. There are two central phenomena the lexical analysis of DGS signs has to cope with: iconicity and mouthing.

4.1 Iconicity

Many signs are iconic. The iconic value of a sign can be helpful to determine and differentiate sign lexemes. The underlying image of an iconic sign may in many cases be interpreted "literally" as a picture or displayed action. This "visual" interpretation of the sign often reveals one of its conventional core meanings. Different meanings of the same sign are related to each other in some way. They either can be related to the underlying image or they are derived from each other by metonymic or metaphoric processes.

As a consequence, a sign form with different meanings, which are related to each other by the underlying image of the sign, is considered one polysemous lexeme. In the sign entry the meanings are listed as shown in figure 1 (above).

For many iconic signs the underlying image can be re-activated and changed to produce a modified form. This intentional change of form often results in a more specific meaning. Similar sign forms that can be related to a conventional sign on basis of a change of the underlying image with a predictable meaning change are considered modifications of the respective conventional sign. Modifications are dependent sign forms (word forms) of a basic sign. In the dictionary modifications are listed in separate, but related entries. In the electronic version of the dictionary modifications and basic sign are cross-linked.

Signs, which differ slightly in form but are used to express the same meaning and share the same underlying image and visualisation technique, are interpreted as variant forms of each other. For example the sign POWDER1B (see figure 1) has a variant form with a different handshape (thumb touches all other fingers).

4.2 Mouthing

Mouthings are not considered to be part of the sign lexeme. They refer to words of the spoken language, a different language system with other symbols (cf. Ebbinghaus & Hessmann 2001). Mouthings copy or specify the meaning of a sign and therefore have to be taken into account for determining the meaning of a sign. As a consequence, a sign covers different meanings when it is accompanied by different mouthings, i.e. it is polysemous. This is especially true for semantically underspecified iconic signs that allow for a wide range of different but related meanings.

A distinction is to be made with regard to the frequency of mouthing-sign combinations. Some mouthings are frequently used with a conventional sign. Other mouthings are added spontaneously to convey a specific meaning in a given context. In general, these mouthings are in accordance with the underlying image of the sign or with its core meaning.

We call stable mouthing-sign combinations conventional uses of a sign. Meanings consistently associated with a conventional sign are listed in the sign entry. We call spontaneous mouthing-sign combinations productive uses of a sign. These meanings are not considered conventional meanings and are therefore not listed as meanings in the sign entry.

5. Open questions

Even though most assumptions and decisions have proven valuable over time, some questions remain to be answered. One major problem is how to determine the degree of conventionalisation of signs. With regard to sign strings that follow the structure of a German compound, this is an even more complicated question. As of yet, we have no means to determine whether these are instances of lexicalised forms or just ad-hoc combinations of signs with the primary function of providing an adequate context for the mouthed word to facilitate lip-reading. Are there other criteria than frequency (statistical methods) to identify lexicalised sign combinations as true sequential compounds? Are there constraints for the combination of signs such as signs from different regions or signs whose underlying images do not fit the intended meaning? We expect that a larger corpus of natural DGS data can give clues to answer these questions.

Since in many vocational fields there are no or few deaf experts working together and communicating in sign language, it is hard to imagine, how decisions about conventionalised "technical signs" as equivalents to established technical terms in spoken language can be based on empirical data of natural signing. One lesson we learned from the evaluation of elicited and conversational data is that there are few conventionalised technical signs and a large variety of DGS equivalents. We decided to show these differences and to give insight into general sign formation processes.

6. References

Arbeitsgruppe Fachgebärdenlexika (Ed.) (1996). *Fachgebärdenlexikon Psychologie*. Hamburg: Signum. URL: http://www.sign-lang.uni-hamburg.de/plex (last accessed March 2008).

Arbeitsgruppe Fachgebärdenlexika (Ed.) (1998). *Fachgebärdenlexikon Tischler/Schreiner*. Hamburg: Signum. URL: http://www.sign-lang.uni-hamburg.de/tlex (last accessed March 2008).

Ebbinghaus, H., Hessmann, J. (2001). Sign language as multidimensional communication: Why manual signs, mouthings, and mouth gestures are three different things. In P. Boyes Braem, R. Sutton-Spence (Eds.), *The hands are the head of the mouth: the mouth as articulator in sign language*. Hamburg: Signum, pp. 133--151.

König, S. Konrad, R. Langer, G. (in preparation). What's in a sign? Theoretical lessons from practical sign language lexicography. In J. Quer (Ed.), *Signs of the time. Selected papers from TISLR 2004*. Hamburg: Signum.

Konrad, R., Hanke, T., Schwarz, A., Prillwitz, S., Bentele, S. (2000). *Fachgebärdenlexikon Hauswirtschaft*. Hamburg: Signum. URL: http:// www.sign-lang.uni-hamburg.de/hlex (last accessed March 2008).

Konrad, R., Schwarz, A., König, S., Langer, G., Hanke, T., Prillwitz, S. (2003). *Fachgebärdenlexikon Sozialarbeit/Sozialpädagogik*. Hamburg: Signum. URL: http://www.sign-lang.uni-hamburg.de/slex (last accessed March 2008).

Konrad, R., Langer, G., König, S., Schwarz, A., Hanke, T., Prillwitz, S. (Ed.) (2007). *Fachgebärdenlexikon Gesundheit und Pflege*. Seedorf: Signum. URL: http://www.sign-lang.uni-hamburg.de/glex (last accessed March 2008).

Langer, G., Bentele, S., Konrad, R. (2002). Entdecke die Möglichkeiten. Zum Verhältnis von Mundbild und Gebärde in Bezug auf die Bedeutung in der DGS. *Das Zeichen* 59, pp. 84--97.

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