iLex – A tool for Sign Language Lexicography and Corpus Analysis

Thomas Hanke

University of Hamburg
Institute of German Sign Language and Communication of the Deaf
Binderstraße 34, 20146 Hamburg, Germany
Thomas.Hanke@sign-lang.uni-hamburg.de

Abstract

This paper describes a tool that combines features found in empirical sign language lexicography and in sign language discourse transcription. It supports the user in lexicon building while working on the transcription of a corpus. While it tries to reach a certain level of compatibility with upcoming multimedia annotation tools, it offers a number of unique features considered essential due to the specific nature of sign languages.

1. Introduction

Sign languages are the preferred communications medium for most Deaf people around the world. Sign language uses a number of visually distinctively recognisable articulators (hands, facial expression, mouth, body) in parallel and fully exploits spatial and temporal relations to establish grammatical features.

It is therefore not surprising that sign language researchers had been among the first to integrate digital video into tools for corpus analysis and lexicographic work. However, solutions general enough to cover a broad range of research questions still do not exist. The approach presented here integrates corpus transcription with tools previously used in empirical lexicography.

In the long run, we expect upcoming multimedia annotation tools to include most of the features we consider essential, allowing sign language researchers to use the same tools as the community at large. So this paper can also be viewed as a wish list of features we would like to see in those tools.

2. Sign language notation

The fact that sign languages have no widespread writing systems has major implications for the field: To label and identify signs, most researchers use glosses, i.e. spoken language words semantically overlapping to a large extent with the sign to be identified, relying on the reader's knowledge of the target language. While convenient, this method has definitive weaknesses. It becomes clear that easy access to the original data is essential in such a context.

Before the advent of digital video, phonetic notation served as a substitute to the original data as the video on tape was too cumbersome to be checked. Although production of phonetic transcription is an extremely time-consuming task due to the complexity of sign languages' phonological and morphological structure, notation does not become superfluous where access to the original data is readily available by way of digital video as it makes data searchable for phonetic aspects (cf. Hanke/Prillwitz 1995 and Hanke, 2001).

The notation system we use, HamNoSys (Hamburg Notation System; cf. Prillwitz et al., 1989 and Schmaling/Hanke, 2001) is alphabetic and non-roman. It describes the parameters handshape, hand orientation, initial location, and dynamics (movements) at a granularity that is believed to be sufficient from a phonological perspective. Although the glyphs for most symbols have been designed along iconicity principles, the relatively large number of characters (about 220) and a rather complex syntax for a phonetic description (due to the fact that both simultaneity and sequentiality need to be described) require input support. Our system currently implements inline syntax-checking and a virtual keyboard.

HamNoSys notations can either be written as strings of characters or in a multi-tier representation to split between the dominant and the non-dominant hand as well as non-manual articulators. It is therefore well-suited both for lexicographic work and transcription.

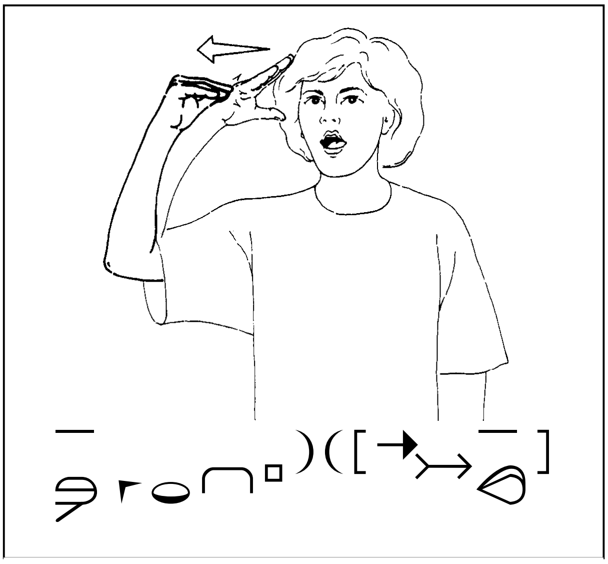


Figure 1. A sign that is glossed HAMBURG1B in our database together with its HamNoSys notation

A HamNoSys-compatible XML application, SiGML (cf. Elliott et al., 2001), has been defined that also forms the bridge to an animation generator that helps in verifying the notation by comparing its animation with the original data on digital video.

In transcription tasks with emphasis on non-manual behaviour, the HamNoSys approach to apply the same set of movement operators to the hands and other parts of the body does not deliver enough detail, e.g. with respect to facial expressions and especially mouthing activities. For these cases, value sets for a number of features have been defined that are included in SiGML as well.

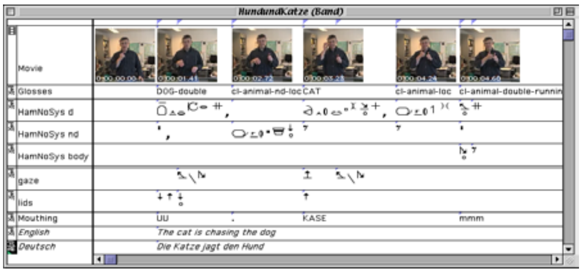
```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE sigml SYSTEM
 "http://www.visicast.org/sigml/SiGML_h4_10.dtd">
<sigml>

<hamgestural_sign gloss="HAMBURG1B">
  <sign_manual>
    <handconfig handshape="ceeall"
            mainbend="bent" ceeopening="slack"/>
    <handconfig extfidir="ul"/>
    <handconfig palmor="d"/>
    <location_bodyarm location="forehead"
         side="right_beside" contact="close"/>
    <par_motion>
      <directedmotion direction="r"/>
      <tgt_motion>
        <changeposture/>
        <handconfig handshape="pinchall"
                           mainbend="bent"/>
      </tgt_motion>
    </par_motion>
  </sign_manual>
</hamgestural_sign>

</sigml>
```

Figure 2. The SiGML notation for the sign shown in figure 1

3. Corpus transcription

Typologically, the sign languages researched so far fall into the category of polycomponential languages. However, morphemes cannot only be uttered sequentially, as in spoken languages, but also co-temporally. It is therefore essential to handle multi-tier representations with lots of tiers per signer involved. Timing granularities needed range from above to below sign (word) level.

Most corpus projects at our institute (child language as well as longitudinal studies on adult signers) up to now used syncWRITER, an interlinear text editor co-developed at Hamburg University from 1989 to 1992 (cf. Papaspyrou/Zienert, 1991; Hanke/Prillwitz, 1995; Hanke, 2001). Its main focus is on entering and easily revising data that can then be rendered into a presentable score/partitur format that can be included in scientific papers and similar applications. Atoms in different tiers can be freely synchronised as long this does not create a cyclic structure.



Glosses: DOG-double cl-animal-nd-loc CAT cl-animal-loc cl-animal-double-runnin

Mouthing: UU KASE mmm

English: The cat is chasing the dog

Deutsch: Die Katze jagt den Hund

Figure 3. syncWRITER's Tracks window: Tiers extending to the right without linebreaks, segments take as much space as required by their textual representation

syncWRITER's major advantage is the seamless integration with video supporting the transcriber's work. Analysis functions, on the other hand, are relatively basic. Anything beyond searching and counting is left to the user to be scripted. It turned out, however, that the built-in support for the AppleScript Object Model is not application-oriented enough to be used by most researchers.

Via the scripting engine, it is possible to link the syncWRITER document to a database, e.g. to look up citation forms of signs indexed by the gloss entered into the transcript. However, as this mechanism is put on top of the programme instead of being an integral part, is not possible to enforce integrity between the transcript and the look-up database.

With a document-based approach, syncWRITER is not really well-suited for teamwork. In addition, the synchronisation mechanism, while being perfectly suited for interlinear text and discourse presentation tasks, is a substantial drawback for analysis tasks as it synchronises points of time instead of time intervals.

4. Lexicographic tools

For the last twelve years, we have been working on a series of specialist sign language dictionaries (computer technology, linguistics, psychology, carpentry, home economics, and social services). With the third dictionary, we began using an empirical approach, i.e. the dictionaries are based on a corpus. Signs in the semantic field of each dictionary are collected from deaf specialists by elicitation and guided interviews. A list of concepts to be covered is driving the selection process. These might be realised in sign language as simple signs, compound signs, or phrases.

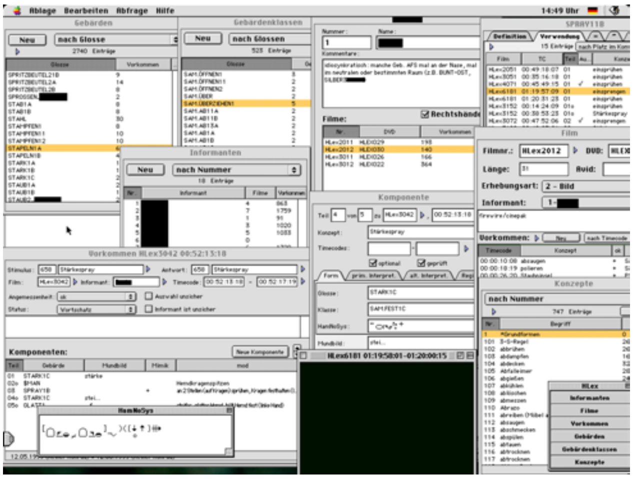
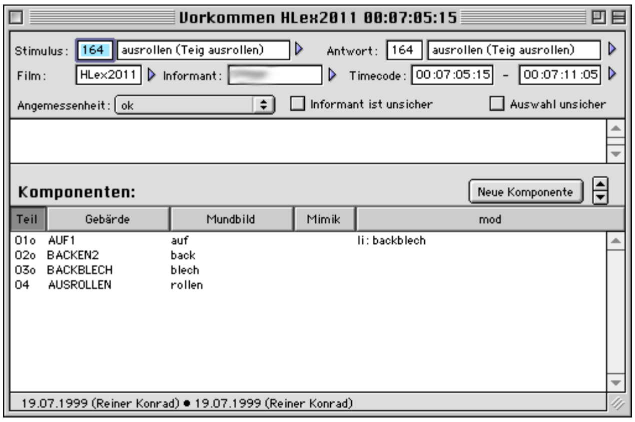


Figure 4. Transcriber's working environment with multiple GlossLexer windows. Informant-specific data is blackened

GlossLexer (Hanke et al., 2001), the multimedia lexical database tool used in these projects, allows the user to transcribe the elicited replies ("occurrences") on a sign-by-sign level with a fixed set of tiers. Synchronisation on a sub-sign level is not supported. Values in the master tier are references to the sign table of the relational database; values in the other tiers are unrestricted text. Tags are dense within an occurrence by default, i.e. the tagged time intervals overlap at their endpoints. This can be changed to leave gaps since pauses as well as off-topic conversation is left uncoded in these projects.



Vorkommen HLex2011 00:07:05:15

Stimulus: 164 ausrollen (Teig ausrollen) Antwort: 164 ausrollen (Teig ausrollen)
Film: HLex2011 Informant: [redacted] Timecode: 00:07:05:15 - 00:07:11:05
Angemessenheit: ok □ Informant ist unsicher □ Auswahl unsicher

Komponenten: Neue Komponente

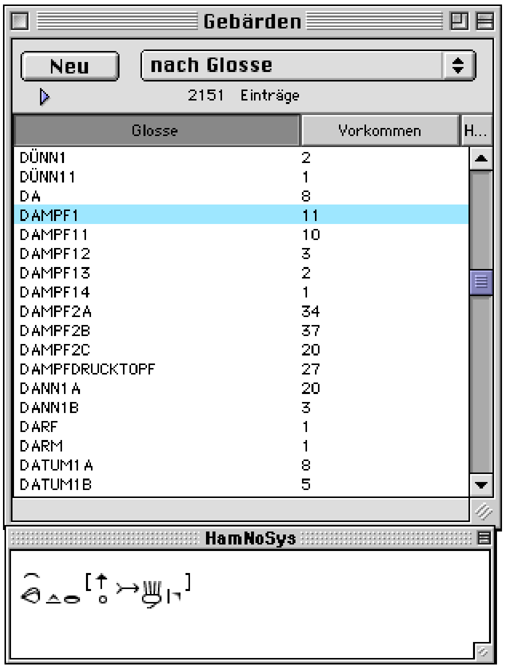
Teil	Gebärde	Mundbild	Mimik	mod
01o	AUF1	auf		li: backblech
02o	BACKEN2	back		
03o	BACKBLECH	blech		
04	AUSROLLEN	rollen		

19.07.1999 (Reiner Konrad) ● 19.07.1999 (Reiner Konrad)

Figure 5. Occurrence consisting of a sequence of four signs

When having to assign the observed token to a type, the transcriber can review either the citation forms or all occurrences of the signs in question by looking at their notation and/or digital video.

Iconicity is one of the key features of sign languages. This does not mean, however, that the semantics of most signs can be derived from the form. Instead, it makes sense to work with a two-level description of lexical items: On the first level, we have the types, i.e. form-meaning pairs, as abstractions from the tokens in the data, very much comparable to lexical items in spoken language. On the second level, form-meaning pairs sharing the same underlying image are mapped onto one type. This level of analysis is very helpful in analysing the widespread productive use of signs. On both levels of description, form-related relations such as homophony as well as semantic relations can be maintained. (Due to deficiencies in the underlying notation system HamNoSys – mirroring the state of the art in that field – even homophony is only suggested by the system and can be overridden by the user.



Gebärden

Neu nach Glosse

2151 Einträge

Glosse	Vorkommen	H...
DÜNN1	2	
DÜNN11	1	
DA	8	
DAMPF1	11	
DAMPF11	10	
DAMPF12	3	
DAMPF13	2	
DAMPF14	1	
DAMPF2A	34	
DAMPF2B	37	
DAMPF2C	20	
DAMPFDRUCKTOPF	27	
DANN1A	20	
DANN1B	3	
DARF	1	
DARM	1	
DATUM1A	8	
DATUM1B	5	

HamNoSys

Figure 6. List of available signs. The bottom window shows HamNoSys for the selected sign. Clicking on the triangle plays the citation form for that sign.

5. Integrating the two approaches

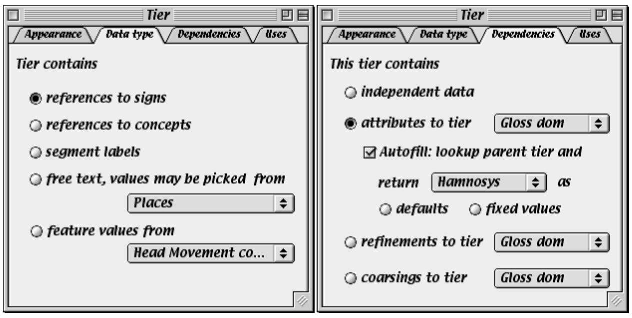
For building larger sign language corpora, the maintenance of a corpus lexicon is absolutely necessary. This is a direct consequence of the missing writing system and the failure of current notation system to provide an orthography. For smaller projects, this lexicon can be substituted by the transcriber's intimate knowledge of the data, but this is not case for large projects that typically are team efforts. In our case that meant that the project had to decide between detailed transcription as possible in syncWRITER and full lexicon support provided by GlossLexer while in fact they needed a combination of both.

In 2000, we decided to develop the necessary tool as none of the tools available in the LR community fulfilled our requirements list. For resources reasons, we opted for extending GlossLexer to allow detailed transcription[1]. As a consequence, the resulting tool, iLex, has a definite bias towards lexicon building. Currently, the sophistication of the transcription user interface does not match what researchers were used to from working with syncWRITER.

The major extension is the introduction of tiers and tags as tables into the relational database. Depending on the type of tier they are assigned to, tags contain free-format text, or references to other database tables, such as signs. For all classes of restricted input, the user has a browser available to choose among the possible values. By this construction, the database automatically guarantees data integrity not only for the lexicon, but also for the corpus and the lexicon together.

Dependencies between tiers may be defined, resulting in aligned structures. As the number of independent tiers per person in the dialogue is not restricted to just one, fine-grained annotation not resulting in aligned structures is still achievable.

For the user's convenience, tiers can be grouped in



Tier

Appearance / Data type / Dependencies / Uses

Tier contains

● references to signs
○ references to concepts
○ segment labels
○ free text, values may be picked from Places
○ feature values from Head Movement co...

Tier

Appearance / Data type / Dependencies / Uses

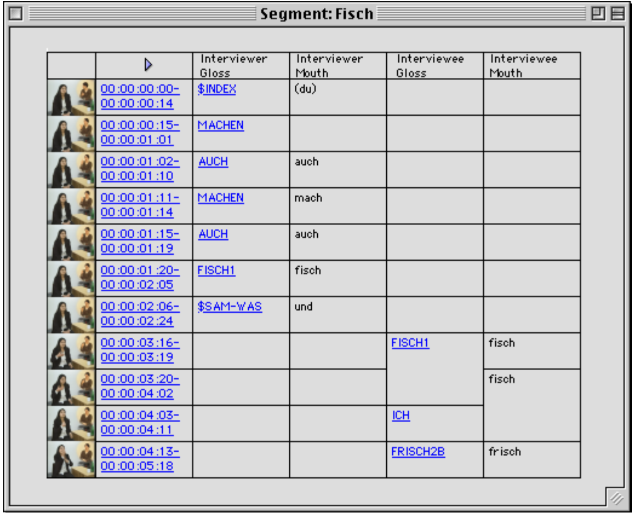
This tier contains

○ independent data
● attributes to tier Gloss dom
☑ Autofill: lookup parent tier and return Hamnosys as
○ defaults ○ fixed values
○ refinements to tier Gloss dom
○ coarsings to tier Gloss dom

Figure 7. Second and third panel of the Tier definition

schemes (such as “interviewee” or “interviewer”). Thus a set of tiers, together with their visibility rules as well as font and size defaults, can be added at once.

Multi-tier data are represented on-screen either in table format with a top-down flow of time or aligned to a linear time line. In the latter case, the user can switch between a top-down and a left-right flow of time. The interlinear text representation as featured by syncWRITER where each event takes up as much space as required by its textual representation has not been implemented. However, the score format (except video thumbnails) can be produced by exporting the tag structure in XML format and using EXMARaLDA (Schmidt, 2001) to produce an RTF (Rich Text Format) score document.



Segment: Fisch

	▷	Interviewer Gloss	Interviewer Mouth	Interviewee Gloss	Interviewee Mouth
	00:00:00:00- 00:00:00:14	$INDEX	(du)		
	00:00:00:15- 00:00:01:01	MACHEN			
	00:00:01:02- 00:00:01:10	AUCH	auch		
	00:00:01:11- 00:00:01:14	MACHEN	mach		
	00:00:01:15- 00:00:01:19	AUCH	auch		
	00:00:01:20- 00:00:02:05	FISCH1	fisch		
	00:00:02:06- 00:00:02:24	$SAM-WAS	und		
	00:00:03:16- 00:00:03:19			FISCH1	fisch
	00:00:03:20- 00:00:04:02				fisch
	00:00:04:03- 00:00:04:11			ICH	
	00:00:04:13- 00:00:05:18			FRISCH2B	frisch

Figure 8. Multiple tiers in table format

iLex has been implemented on top of a commercial object-oriented framework and runs on MacOS, MacOSX, and Windows 2000. For displaying video, it relies on Apple QuickTime. The SQL database is connected via ODBC.

6. Outlook

While iLex enables transcription of signed text with direct access to a lexicon, search facilities currently completely rely on SQL to query the relational database. It is well-known fact that time-related structures as those in corpus annotation are difficult and inefficient to express in SQL. It is therefore desirable to either integrate special-purpose search engines or to transfer to a system that has such a facility, once environments are open enough to allow for the sign language specific features to be added.

Another deficit is missing support for co-references. While this has recently been added to HamNoSys (Schmaling/Hanke, 2001), iLex has no knowledge of the textual descriptions for co-references and therefore cannot maintain integrity.

Support for crosslinguistic research, especially in language mixing and switching situations, has only recently been added and is an ad-hoc solution. We have to await experiences with this special kind of annotation work to see which further developments are necessary.

In the long run, it will also be useful to join the empirical lexicon structures with those currently under development for an HPSG-based machine-translation dictionary for sign languages (Hanke, 2002).

7. References

Elliott, R., J.R.W. Glauert, R. Kennaway and I. Marshall, 2000. The development of language processing support for the ViSiCAST project. In *The fourth international ACM conference on Assistive technologies ASSETS*. New York: ACM. 101–108.

Hanke, T., 2001. Sign language transcription with syncWRITER. *Sign Language and Linguistics* 4(1/2): 267–275.

Hanke, T., 2002. HamNoSys in a sign language generation context. In R. Schulmeister and H. Reinitzer (eds.), *Progress in sign language research: in honor of Siegmund Prillwitz / Fortschritte in der Gebärdensprachforschung: Festschrift für Siegmund Prillwitz*. Seedorf: Signum. 249–266.

Hanke, T., R. Konrad and A. Schwarz, 2001. GlossLexer – A multimedia lexical database for sign language dictionary compilation. *Sign Language and Linguistics* 4(1/2): 161–179.

Hanke, T. and S. Prillwitz. 1995. syncWRITER: Integrating video into the transcription and analysis of sign language. In H. Bos and T. Schermer (eds.), *Sign Language Research* 1994*: Proceedings of the Fourth European Congress on Sign Language Research, Munich, September 1-3, 1994*. Hamburg: Signum. 303–312.

Johnston, T., 1991. Transcription and glossing of sign language texts: Examples from AUSLAN (Australian Sign Language). *International Journal of Sign Linguistics* 2(1): 3–28.

Papaspyrou, C. and H. Zienert, 1991. The syncWRITER programme. In S. Prillwitz and T. Vollhaber (eds.), *Sign language research and application: Proceedings of the international congress on sign language research and application, March 23-25, 1990 in Hamburg*. Hamburg: Signum. 275–294.

Prillwitz, S. et al., 1989. *HamNoSys. Version 2.0; Hamburg Notation System for Sign Languages. An introductory guide*. Hamburg: Signum.

Schmaling, C. and T. Hanke, 2001. HamNoSys 4.0. In: T. Hanke (ed.), *Interface definitions*. ViSiCAST Deliverable D5-1. [This chapter is available at http://www.sign-lang.uni-hamburg.de/projekte/HamNoSys/HNS4.0/englisch/HNS4.pdf]

Schmidt, T., 2001. The transcription system EXMARaLDA: An application of the annotation graph formalism as the basis of a database of multilingual spoken discourse. In S. Bird, P. Buneman and M. Liberman (eds.), *Proceedings of the IRCS Workshop On Linguistic Databases, 11-13 December 2001*. Philadelphia: University of Pennsylvania. 219–227.

[1] From a theoretical point of view this is rather unexpected as lexicographic work based on empirical data can easily be subsumed under corpus annotation.