iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography

Thomas Hanke, Jakob Storz
Institute of German Sign Language and Communication of the Deaf
University of Hamburg
Binderstraße 34, 20146 Hamburg, Germany
E-mail: {thomas.hanke,jakob.storz}@sign-lang.uni-hamburg.de

Abstract

This paper presents iLex, a software tool targeted at both corpus linguistics and lexicography. It is now a shared belief in the LR community that lexicographic work on any language should be based on a corpus. Conversely, lemmatisation of a sign language corpus requires a lexicon to be built up in parallel. We introduce the basic concepts for transcription work in iLex, especially the interplay between transcripts and the lexicon.

1. Background

For empirical sign language research, the availability of Language Resources, their quality as well as the efficiency of software tools to create new resources is a pressing demand. The software solution iLex is our approach to meet these requirements at least to a certain extent: It is a database system to make existing resources available, and it is a tool to create new resources and to manage their quality.

Language resources for sign languages are special insofar as there is no established writing system for any sign language in the world. Notation systems can only partially fill this gap, and their most important drawback is the effort needed to describe signed utterances in enough detail that would allow the researcher to do without going back to the original data. In the early 1990ies, syncWRITER (Hanke & Prillwitz, 1995; Hanke, 2001) was our first attempt for a transcription tool that not only allowed the user to link digital video sequences to specific parts of the transcription, but also allowed the video to become the skeleton of the transcription. The drawback of that solution was that it was mainly targeted towards the presentation of the transcriptions in a graphically appealing way, but was not equally well equipped for any discourse-analytic or lexicographic purpose.

In the context of a series of special terminology dictionaries, we therefore developed an independent tool, GlossLexer (Hanke et al., 2001), concentrating on the development and production of sign language dictionaries, both in print and as multimedia hypertexts, derived from transcriptions of elicited sign language utterances. At the heart of this tool was a lexical database, growing with the transcriptions. This tool, however, was not suitable to adequately describe really complex signed utterances, as it reduced them to sequences of lexical entities as suitable only in a purely lexicographic approach.

iLex (short for "integrated lexicon", cf. Hanke, 2002b) now combines the two approaches: It is a transcription database for sign language in all its complexity combined with a lexical database. In iLex, transcriptions do not consist of sequences of glosses typed in and time-aligned to the video. Instead, transcriptions consist of tokens, i.e. exemplars of occurrences of types (signs) referencing their respective types. This has immediate relevance for the lemmatisation process. Due to the lack of a writing system, this is not a relatively straightforward process as for spoken languages with a written form featuring an orthography, but requires the transcriber's full attention in type-token matching.

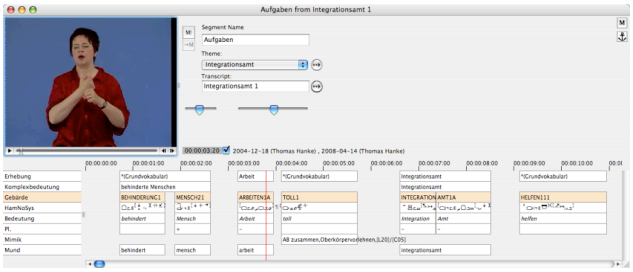
By providing tool support for this process, iLex enables larger and multi-person projects to create transcriptions with quality measures including intra-transcriber and inter-transcriber consistency.

For a research institute as a whole, the central multi-user database approach means that all data are available at well-defined places, avoiding data loss often occurring in a document-centric approach as researchers and students leave and enabling an effective data archiving strategy. Finally, combining data from several projects often is the key to achieve the "critical mass" for LR-specific research.

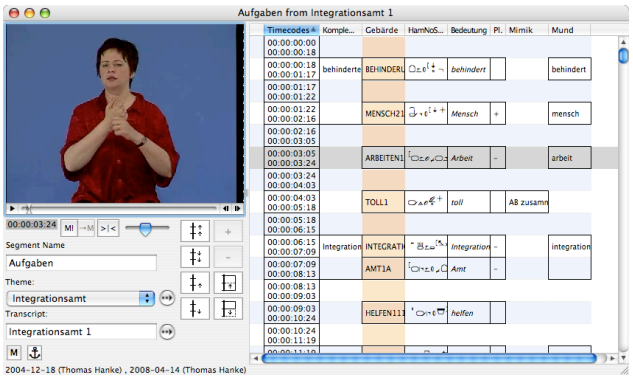
At the IDGS in Hamburg, iLex today is not only used in discourse analysis and lexicography, but a number of applied areas draw from the data collected and contribute themselves: The avatar projects ViSiCAST and eSIGN allow transcripts from the database to be played back by virtual signers (Hanke, 2002a; Hanke, 2004a); in computer-assisted language learning for sign languages, authoring tools can directly import iLex transcripts (Hanke, 2006).

2. Flow of Time

iLex features a horizontal view of transcript data familiar to those using any other transcription environment: Time flows from left to right, and the length of a tag is proportional to its duration.



This view is complemented by a vertical view, where time flows from top to bottom. Each smallest interval of interest here occupies one row, irrespective of its length. A tag spans one or more such intervals. Unless it is partially overlapping with other tags, the tag is identical to one interval. The focus here is on interesting parts of the transcription, not on the flow of time. If the transcriber detects that two events are not fully cotemporal, but that one starts slightly after the other, for example, the time interval that the two tags have shared so far is split at the point of time where the second event really starts, and the second tag's starting point is moved down one line. This procedure ensures that slightly deviating interval boundaries are possible, but only as a result of a deliberate action by the user.

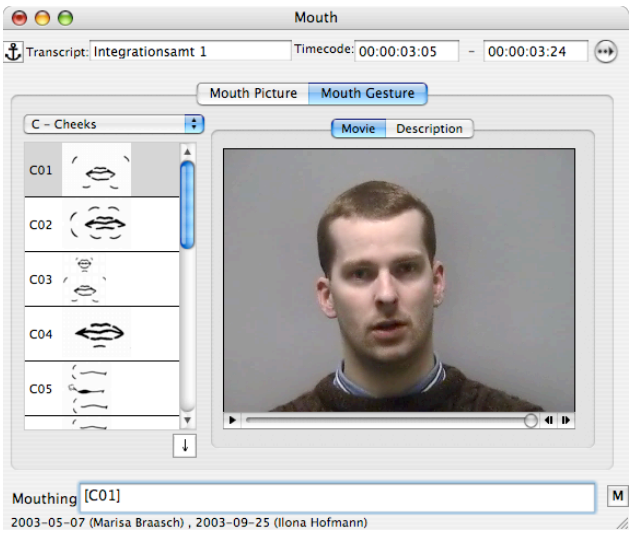


Which of these two views is used is determined by the current task, but also the user's preference. In any case, switching to the other view sheds new light on the transcription and thereby helps to spot errors.

3. A Data Model for Transcripts

Despite the fact that iLex is the only transcription tool used in sign language research with a database instead of a document-centric approach, the data model for transcripts is more or less shared with other tools[1]: Transcripts are linked to a video[2] and have any number of tiers; a tier contains tags that are time-aligned to the video. Tier-to-tier relations define restrictions on the alignment of tags with respect to tags in superordinate tiers. However, iLex goes beyond this by introducing different kinds of tiers. The most important kinds are:

- *Token tiers* contain tokens as tags, i.e. they describe individual occurrences of signs and as such are the most important part of a transcription. iLex allows double-handed and two-handed tokens, or partially overlapping one-handed tokens, but always ensures that the tokens at any point of time do not describe more than two hands per informant.
- In elicitation settings, *answer tiers* group tokens that are signed in response to a specific elicitation, describing the elicitation by referring to a picture, movie segment or text.
- Tags in *phrase structure tiers* group tokens into constituents or multi-sign expressions.
- Tags in *text tiers* simply have text labels. This is the kind of tags found in most other transcription environments. iLex allows the user to assign vocabularies to tiers, so that tags can be chosen from pre-defined lists of values. User-defined vocabularies can be open or closed, but iLex also offers a number of built-in vocabularies with special editors, e.g. in order to tag mouth gestures.



- Tags in *numerical data tiers* can be linked to horizontal and vertical coordinates in the movie frame. Thus, the user can enter data for these tags by clicking into the movie frame, e.g. to track the position of the eye or to measure distances. Tags could also be automatically created by external image processing routines indicating e.g. a likelihood for certain types of events, as a first step to semi-automatic annotation.
- Tags in *value (computed) tiers* are automatically inserted by the system as the user enters data into other tiers. E.g. a tier can be set up to show the citation form of the types referenced by tokens in another tier, in our case by means of a HamNoSys notation (Hanke, 2004b).

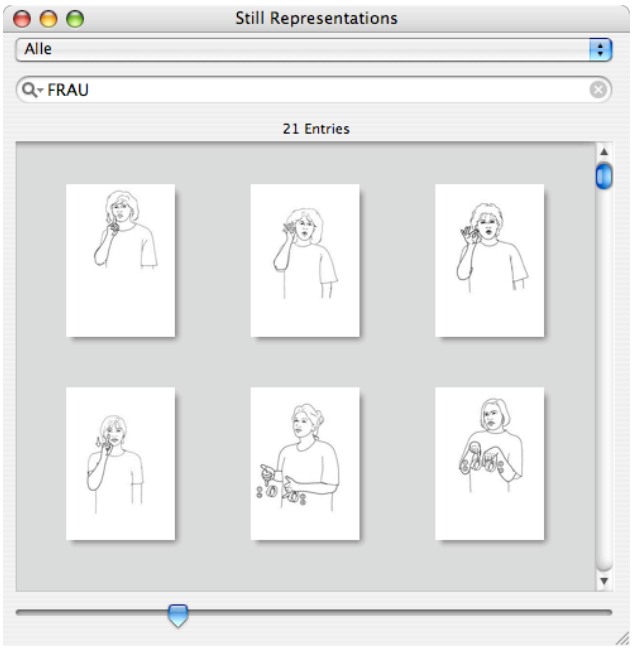
As with most database entities in iLex, the user can easily add metadata to transcripts, tiers, and tags. These may be ad-hoc comments, markers for later review, judgements, or structured data as defined by the IMDI metadata set or its extension for sign language transcription (cf. Crasborn & Hanke, 2003).

[1] As the other systems, iLex's data model can be considered an implementation of the annotation model developed by Bird and Liberman (2001).

[2] iLex transcripts can link to only one "movie". This is no restriction, as iLex works well with movies containing more than one video track. At any point of time, the user can choose to hide tracks s/he is not currently interested in, e.g. close-up views that will only be used in mouthing or facial movements analysis.

4. Lemmatisation

Type-token matching is at the heart of transcribing with iLex, and iLex supports the user in this task. The user can identify candidates for the type to be related to a token by (partial) glosses, (partial) form descriptions in HamNoSys or meaning attributions. The search can be narrowed down by browsing through the types found, comparing tokens already assigned to a type with the token in question. By using alternatives such as browsing tokens or stills, an active competence in HamNoSys (or another notation system used in iLex instead) is not necessary.



Once the right type has been identified, it can easily be dragged into the transcript to establish the token. This procedure avoids simple errors such as typos, and allows for easy repairs. If it is later decided that a type needs to be split into several as form variation seems not to be free, tokens can be reviewed and reassigned (i.e. dragged into the new type) as necessary.

In the token, iLex used to provide a text to describe how the actual instance of the sign deviated from the citation form. The latest version categorises modifications in order to further reduce inconsistent labelling in this part as well.

5. Importing Data from other Transcription Systems

Importing transcripts from other sources, such as ELAN, syncWRITER or SignStream documents (cf. Crasborn et al., 2004; Neidle, 2001), is done by a simple menu command. The results of this import process, however, are transcripts with only text tiers, and a second step is necessary to convert the text tiers describing tokens (in most cases by means of glosses) to real token tiers. iLex supports this process by learning a source-specific mapping table from external glosses to types and modifications in iLex. As inconsistencies may occur in the imported data if lemmatisation was not done rigidly, the transcriber's attention is required. More than one name for a single type is easily dealt with in the mapping mechanism. Different types under the same gloss label, on the other hand, require close inspection of each token assigned.

6. Dictionary Production

In the case of our special terminology dictionaries (cf. König et al., this volume), all of the data needed to produce the dictionary are stored in the database as the results of the transcription process or later analysis steps. This allows automatic production of a dictionary within reasonable time. For that, we use Perl as a scripting language linking the database with Adobe Indesign for layouting the print product and an HTML template toolkit to produce web applications. By just changing the templates (or adding another set), we can completely change the appearance of the dictionary and reproduce print and online versions within hours. Currently, we are developing another set of templates to optimise HTML output for iPhone/iPod touch devices that promise to become an ideal delivery platform for our dictionaries.

7. Collaborative Approach

Using a central database for all people working in a project or even several projects at one institution not only serves data sustainability, but also allows for cooperative work. First and foremost, each transcriber contributes to the pool of types as well as tokens for each type making type-token matching easier or at least better informed. Other data, such as project-specific data views or filters, are easily shared between users. The results of introspection can quickly be made available to other users by using a webcam. Integration of camera support into the program allows sharing signed samples without the need to care about technical aspects such as video compression; appropriate metadata for the new video material is automatically added to the database.

The newest version of iLex takes a first step in supporting Web 2.0 technologies for collaboration: All data can be referenced by URLs. By simply dragging data from an iLex window into a Wiki or Blog, the URL is inserted and anyone with access to the iLex database can view the data talked about in a discussion by simply clicking onto the URL.

The "disadvantage" of collaboration of course is the need to agree on certain transcription conventions. While many aspects of the transcription process can be individualised, other data, such as the types inventory, need to be accessed by all users, and therefore need to be understood by all users; extensions need to be made in a consistent manner. Experience shows that a couple of meetings with all transcribers are needed if a new project is set up to work with the pool, especially if the new project's target differ significantly from what the other projects do.

8. Technical Background

The name iLex stands for the transcription database as well as the front-end application used to access it.

The database normally resides on a dedicated or virtual database server. As the SQL database engine, we have chosen PostgreSQL, an open-source database server system that can be installed on a wide variety of

platforms.[3] It is rock-solid and has well-defined security mechanisms built in, it is well supported by an active user community, and features a couple of implementation aspects that are advantageous in our context, such as server-side inclusion of scripting languages including Perl.

Movies, stills and illustrations are not stored in the database, but only references to them. They can either reside on the users' computer or on a central file server. With video archives becoming rather large over time, of course only the second solution is viable in the long run.[4] This hybrid storage concept also allows users to work from home: Access to the database is low-bandwidth and therefore can be secured with a virtual private network approach, whereas the user can locally access the video currently in work without a performance hit. Tokens from other videos not available on the local computer then come over the network, but usually are that short that even slower connections should be fine.

The front-end software is available free of charge for MacOS X as well as Windows XP (with a couple features only available for MacOS), with German and English as user interface languages. (Localisation to other languages is easily possible.) Upon request, source code for the front end is also available except for a couple of functions where we decided to use commercial plug-ins instead of implementing the services ourselves.

For single-user applications, the server and the client can be installed on the same machine, even on a laptop. However, unless that machine has plenty of RAM, page swapping will reduce the processing speed compared to a standard server-client scenario.

[3] At the IDGS, we currently use a dedicated four-cores Mac Pro with 6 GBytes of memory and a mirrored harddisk. At some times, as many as 20 persons access the server without any experiencing any performance reductions.

[4] At the IDGS, we use a dedicated MacOS X Server file server with a storage area network (current size: 8 TB). We have experimented with video streaming servers before, but found that users rarely view more than a couple of seconds of a movie at once. In this situation, the negotiation overhead associated with streaming costs more than the streaming itself saves.

9. References

Bird, S and M. Liberman (2001). A formal framework for linguistic annotation. *Speech Communication* 33(1,2), pp. 131–162.

Crasborn, O. and T. Hanke (2003). *Metadata for sign language corpora*. Available online at: http://www.let.ru.nl/sign-lang/echo/docs/ECHO_Metadata_SL.pdf.

Hanke, T. (2001). Sign language transcription with syncWRITER. *Sign Language and Linguistics*. 4(1/2), pp. 275–283.

Hanke, T. (2002a). 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, pp. 249–266.

Hanke, T., (2002b). iLex - A tool for sign language lexicography and corpus analysis. In: M. González Rodriguez, Manuel and C. Paz Suarez Araujo (eds.): *Proceedings of the third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spain.* Paris: ELRA, pp. 923–926.

Hanke, T. (2004a). *Lexical sign language resources – synergies between empirical work and automatic language generation*. Paper presented at LREC 2004, Lisbon, Portugal.

Hanke, T. (2004b). HamNoSys - Representing sign language data in language resources and language processing contexts. In: O. Streiter and C. Vettori (eds.): *Proceedings of the Workshop on Representing and Processing of Sign Languages, LREC 2004, Lisbon, Portugal*, pp. 1–6.

Hanke, T. (2006). Towards a corpus-based approach to sign language dictionaries. In: C. Vettori (ed.), Proceedings of a Workshop on the representation and processing of sign languages: lexicographic matters and didactic scenarios, LREC 2006, Genova, Italy, pp. 70–73.

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, Germany*. Hamburg: Signum, pp. 303–312.

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), pp. 161–179.

Neidle, C. (2001). SignStream™: A database tool for research on visual-gestural language. *Sign Language and Linguistics*. 4(1/2), pp. 203–214.