iLex: Handling Multi-Camera Recordings

Thomas Hanke, Jakob Storz, Sven Wagner

Institute of German Sign Language and Communication of the Deaf
University of Hamburg, Germany

DGS-KORPUS

INTRODUCTION

More than 15 years ago, we introduced the first sign language transcription environment working with digital video (syncWRITER, cf. Hanke&Prillwitz 1995). However, back then digital video in very small spatial resolution was good enough to show the video in combination with the transcript, but not really to transcribe every detail from it. Rather, one had to use VCRs – either remote-controlled by the transcription environment or directly operated by the transcriber. In the following years, technological advances finally allowed to digitize video full-size SD and then to create digital video directly with the camera and to easily transfer the material to the computer. Now, processing speed and storage capacities would also allow HD videos to be used full-size in a transcription environment. However, even on very large screens, video competes with the space needed for a useful transcription layout. This is even more true so with material that has been shot with multiple cameras. Two of our projects, Dicta-Sign and DGS Corpus, use seven cameras to record a pair of informants, too much to be displayed full-size at the same time.

Sign language transcription environments such as ELAN (Crasborn&Sloetjes 2008) or iLex (Hanke&Storz 2008) have been designed at times when researchers were using digital video in the size of up to half SD (such as 320x240) and certainly need to be improved for the requirements of today's projects delivering multi-camera HD material.

ELAN allows the user to relate several media files to a transcript and to sync them. iLex just allows one single media container and relies on the container format, such as QuickTime, to group and sync several video streams into one container.

To **save screen real estate**, both systems allow the user to vary the display size from a fraction of the videos' spatial resolution to full size (and beyond) for all visible videos. iLex in addition allows the user to switch on or off individual tracks within the media container. This works quite fine with two or three different views grouped, but fails to provide an adequate solution when more camera views are available: A spatial layout of the tracks (defined in the container) that might be optimal when focussing on one informant can be far from optimal in situations where both informants need to be watched in parallel. In both systems, **different display sizes for individual video views** are not possible except by relying heavily on container formats to include one video in multiple sizes and the user switching one on and the others off as needed or to produce copies of one movie in several spatial resolutions.

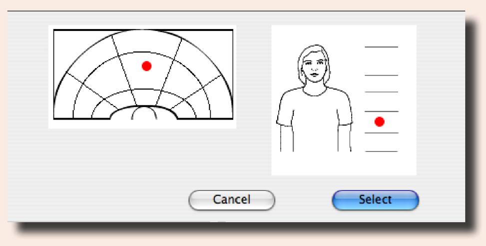
Zooming onto specific parts of a video is also not possible except by providing the zoomed version as a separate movie (cf. Crasborn & Zwitserlood 2008). Here we present a user interface study that promises to deliver the flexibility needed and at the same time to save transfer bandwidth and local processing power which even nowadays are an issue when dealing with several HD videos in parallel.

ILEX IN A NUTSHELL

iLex is a database bringing tother a lexical database and transcription. (Compatibility with other transcription environments is achieved via XML import and export.)

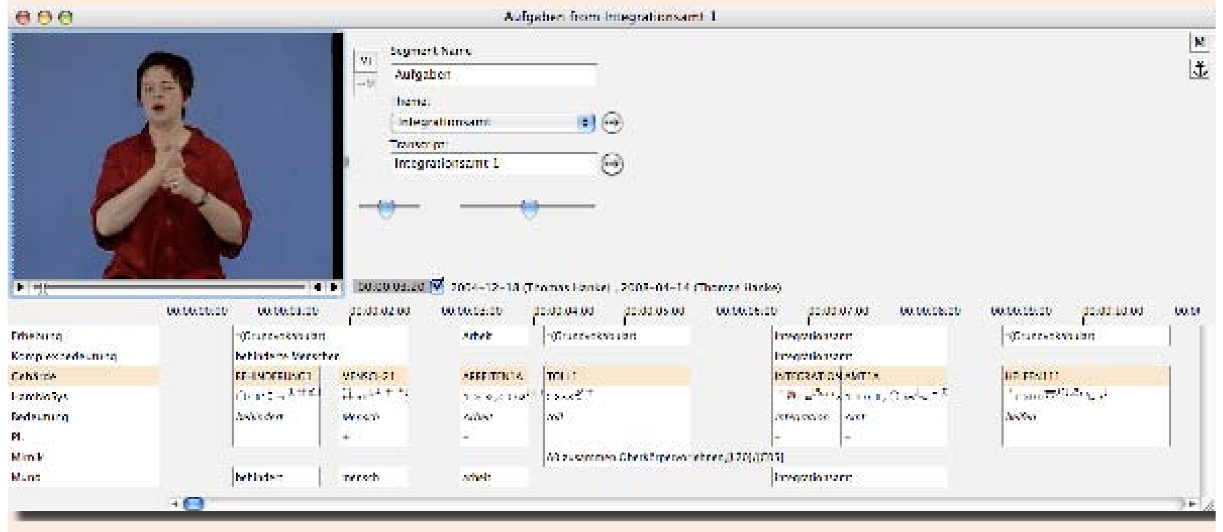
Tags are Database Entities

For many tier types, tags are not text, but rather references to other entities. E.g. token tags refer to the type. This allows direct access from the token to the type to all tokens associated with that type. For other tier types, specialised editors are available.

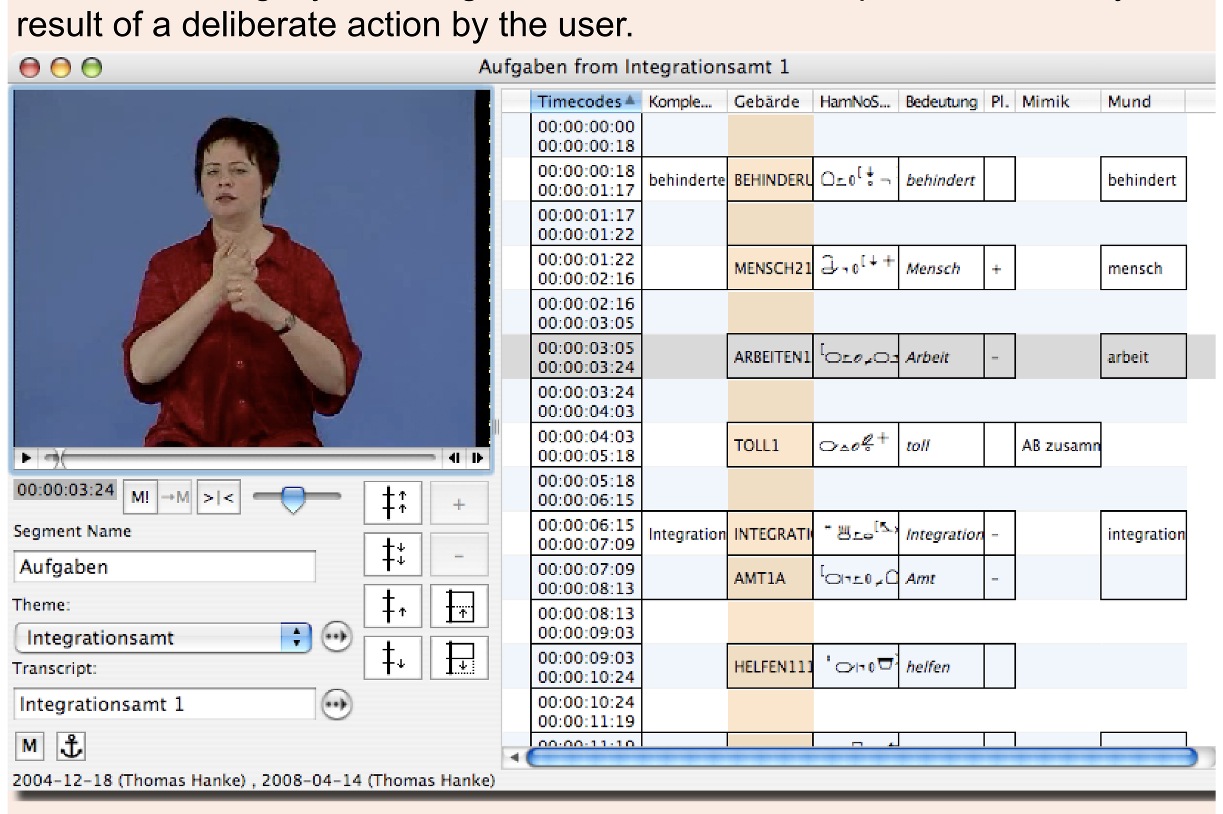


Flow of Time

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.



SCREEN LAYOUTS

In our projects, transcribers have screens with native resolutions of either 1920x1200 or 2560x1440. So except for very rare cases, full HD resolution (1920x1080) is not used for transcription as the movie would occupy a good part of the screen. Depending on what they transcribe, we expect users to work more with 1⁄3 of full HD (640x360), 1⁄4 (480x270) or even 1⁄6 (320x180) rather than with 1⁄2 (960x540). (Users can still resize to any in-between value they prefer. iLex uses the next higher available resolution and scales that down.)

Based on the type of discourse to be described as well as personal preferences, we expect most transcribers to work with one or two movies at a time, optionally with thumb-nail-size view (160x90) for the other cameras.

Focus on one movie at a time

In this layout, clicking on any (movie or still) thumbnail zooms the video shown so far out into a thumbnail and the thumbnail video in to the current large size. When needed, a context menu allows to switch to a two-large-movies layout.

Focus on primary views for both/all informants

With two or more large-size videos shown, thumbnails are bundled to one of the large videos. A click on a thumbnail then exchanges its movie with the bundled one.

Automatic switching based on tagging

Whenever tagging is available that is a good estimator for what the transcriber will need to focus on, this tagging can be used to switch automatically between different layouts. If for example turns have already been tagged, it makes sense to have the signer in a large view and the addressee in a small view. Good approximations to manual turn tagging can hopefully be in the near future achieved automatically through image processing (→Dicta-Sign poster on Sunday). Another source of information is knowledge about the tasks informants are currently working on, as logged by Session Director.

Of course, thumbnail buttons remain available to either switch to secondary views (such as birds-eye views on a single informant) or to the other informant when needed.

Three different multicamera layout examples:
1/6 for B1 and A1 cropped, 1/12 for B2, C, and A1
1/6 for B1 and B2, 1/12 for A1, A2, and C
1/3 for A1 cropped, 1/1 for A1 zoomed to face

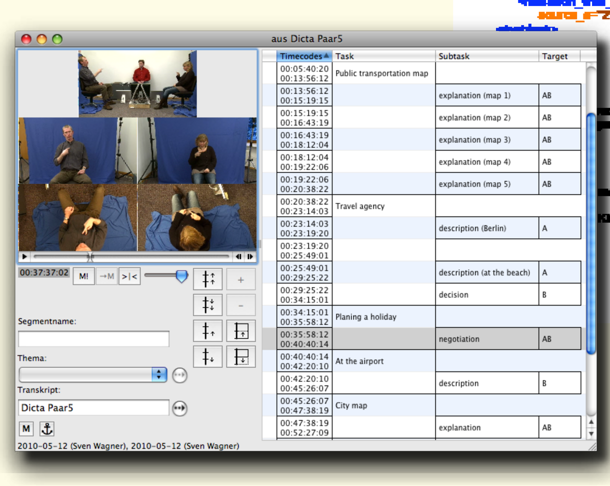
SESSION DIRECTOR DESCRIPTION FILES

In our data collection sessions, we use a custom software, "Session Director", that allows the moderator to present slides to the informants by the click of a button, and to keep track of the time elapsed for each individual task as well as the whole session.

When launched, Session Director loads an XML file describing the session. For each task and subtask, it defines the expected and maximum acceptable duration, the text of the user interface elements visible in Session Director and of course the ids of the slides from an Apple Keynote document to be shown to the informants, either as a common set or separately for each informant. Furthermore, it defines the relative importance of each task which Session Director will eventually use to mark tasks that can be skipped. In addition, text can be entered that will be displayed alongside with the task detail.

In general, the order of tasks, including breaks, is pre-determined by the XML session description. The moderator has, however, the freedom to rearrange tasks, to change the expected duration of the session Session Director measures progress against or to halt the task time with a pause function should a spontaneous break become necessary.

All the moderator's interactions with Session Director are logged with time stamps. This allows us to determine automatically where on the videos certain tasks (or pauses) can be found and also to conclude from our knowledge of the tasks who presumably is the active signer at a given point in time.



DERIVED VIEWS

In addition to the views available through the films actually shot during the data collection, some derived formats are useful for the transcriber. Top of the list with HD sources certainly is **zooming** onto particular parts of the video, such as the signer's face. In the beginning, we ask the user to draw a frame around the signer's face. This may have to be repeated for several points in time in the video, whenever the signer moves significantly. In the future, we hope to automate this windowing through image processing (→poster by Collet et al., on interfaces between transcription environments and image processing). Other examples for derived views include **results of image processing** such as stereo pictures. Changes in spatial or temporal resolution alone are not considered derived views. We try to give the users the impression that any view can be scaled continuously; therefore resolution pyramids are not immediately visible to the user. As we do not see any need at this point of time to work with reductions in temporal resolution (in fact we would like to have higher resolutions available), such reductions are simply not offered as options.

We are still experimenting how to handle **cropping** (cutting away border stripes of the image). The idea with cropping is that anything lying outside the marked area is of no interest for transcription, and therefore the cropped movie could replace the original for all further processing. One of the problems is who might be authorised to apply cropping, as all information outside the cropped area would no longer be visible to any transcriber so errors in cropping might pass undetected.

While results of image processing might not immediately become available to the transcriber, zooms are available to the user at the click of a button: iLex just loads a higher-resolution version of the movie and then lets QuickTime crop the image in memory to the part the user is interested in. If such a derived view is used over a longer period of time, iLex marks this view to be produced as a stand-alone movie to save bandwidth and computing power on the client's side.

VIDEO SERVER INFRASTRUCTURE

Our video server currently consists of three machines with 16 processors each, attached to a SAN with a storage capacity of 100 TBytes. Two thirds of the capacity is reserved for the original footage, one third is available for caching resolution pyramids and other derived video. However, no real caching strategy is in place at this point in time. Instead, cache movies are produced as processing capacity allows. iLex then keeps track of their usage, but purging is currently left to the administrators. Our idea is to observe the system for some time before implementing strategies how to manage cache size. In the current iLex structure which allows the user to copy movies onto the local harddisk in order to work at locations where bandwidth does not allow video server access purging might render local copies useless as iLex would no longer look for them once the database entries are deleted.

Another option for the future is to provide zooming on the server side in real-time. As we currently do this on the client side, we know it can be done in real-time. Implementation on the server side, however, requires much more work, so we will first observe how much this feature will actually be used.

REFERENCES

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd (Eds.), Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages. ELRA, Paris, pp. 39-43.

Crasborn, O. & Zwitserlood, I. (2008). The Corpus NGT: an online corpus for professionals and laymen, In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd (Eds.), Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages. ELRA, Paris, pp. 44-49.

Hanke, T. & Prillwitz, S. (1995). SyncWriter. Integrating Video into the Transcription and Analysis of Sign Language. In T. Schermer & H. Bos (Eds.), Proceedings of the Fourth European Congress on Sign Language Research, Munich, September, 1994. Signum: Hamburg, pp. 303-312.

Hanke, T. & Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd (Eds.), Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages. ELRA, Paris, pp. 64-67.

AKADEMIE DER WISSENSCHAFTEN IN HAMBURG

SEVENTH FRAMEWORK PROGRAMME

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135 and from the German Academies of Science Programme.