Combining video and numeric data in the analysis of sign languages within the ELAN annotation software

Onno Crasborn (Radboud University Nijmegen)

Han Sloetjes, Eric Auer & Peter Wittenburg (Max Planck Institute for Psycholinguistics, Nijmegen)

Abstract

This paper describes hardware and software that can be used for the phonetic study of sign languages. The field of sign language phonetics is characterised, and the hardware that is currently in use is described. The paper focuses on the software that was developed to enable the recording of finger and hand movement data, and the additions to the ELAN annotation software that facilitate the further visualisation and analysis of the data.

1. Introduction

The ELAN software is a linguistic annotation tool that was designed for the creation of text annotations to audio and video files. Starting its life in the domain of speech and gesture research, in recent years it has increasingly been used in the study of sign languages. Aspects of ELAN were enhanced for the creation of the sign language corpus in the EU project European Cultural Heritage Online (ECHO), in which comparable data from three signed languages were collected, annotated, and published (Brugman et al. 2004, Crasborn et al. 2004).

This paper reports on a recent development in ELAN: the integrated display and annotation of numeric data with audio and video data. These numeric data presently stem from one specific movement tracking system that focuses on the hand and fingers, but in principle any type of numeric data can be used. In this paper we focus on the use of such technology in the phonetic analysis of sign languages used by Deaf communities (section 2). We describe the current hardware that is used, as well as the specific nature of the kinematic data generated by the software that we developed (section 3). Section 4 starts with a brief sketch of the ELAN annotation software, and describes in detail what types of functionality have been added to integrate the display and analysis of the kinematic data. We conclude by presenting some possible future software developments that would broaden the use of the current enhancements in ELAN, and discuss ways in which the integration of video and kinematic data can be used for all kinds of exploratory studies in sign linguistics (section 5). However, the focus of the present paper is not on concrete phonetic questions that might be investigated in sign languages, but rather on the methodology of such studies.

2. Sign language phonetics

Just as in spoken languages, sign languages show a distinction between a phonological and a phonetic level of organisation. Lexical items in sign languages typically consist of manual action: short one-handed or two-handed movements. These signs can be characterised in terms of properties like handshape, orientation, location and movement. The lexicon of a sign language consists of a limited number of recurring handshapes, locations, etc. These forms make up the phonological inventory of the language, and have been compared to phonemes or phonological features since Stokoe's (1960) pioneering work (see Brentari 1998 for an overview). In addition, non-manual aspects play a large role in sign language structure, but primarily at the morphosyntactic, pragmatic and discourse levels, and to a much lower degree in the lexicon (see Crasborn 2006 for an overview).

While these phonological specifications are constant for a given lexical item, the phonetic realisation of a sign by a given person at a given moment is highly variable. This phonetic variation has received fairly little attention in the literature (but see Wilbur & Nolen 1986, Wilbur 1990, Wilcox 1992, Cheek 2001, Crasborn 2001). In part, this has been due to the lack of easily accessible tools for measuring the articulation of sign language – the most obvious and accessible level of structure to quantify, and in that respect similar to acoustics in speech. Studies of the phonetic level of sign language could contribute considerably to our linguistic and sociolinguistic knowledge of sign language structure and use.

Phonetic studies that are based on video are limited in their quantitative potential: they are typically used as the basis for transcription, and are restricted in having a fairly low temporal resolution (25Hz for PAL and 29.97Hz for NTSC). The equipment that is described in section 3 allows for the very detailed measurement of hand and finger movements (and in principle any body parts), both in space and time. The aim of the software development described in this paper is to facilitate phonetic studies. The acquisition and use of equipment will remain

problematic given the high costs, but the data analysis will become much easier with the extensions of ELAN described in section 4: the user will be able to integrate the measurements with video recordings, which will facilitate data analysis for phonetic studies.

3. Current equipment for measuring hand and finger movements

3.1. Hardware

The lab setup includes one right hand Virtual Technologies *CyberGlove* with 22 bend sensors and two Ascension *Flock of Bird* electromagnetic location trackers. The glove also has a button and a LED light on it. Those can be used for generic user interface purposes or to place or visualize time markers. One tracker is fixed to the glove, the other tracker can be attached to the other hand or to other reference points on the subject.

The CyberGlove equipment consists of a glove with proprietary resistive bend sensors and a box with electronics. The box is connected to the PC via a RS232 serial port. The two location trackers work with one common transmitter antenna cube. Each tracker has its own electronics, communicating with the other through a fast serial RS485 link. Only one of the tracker electronics boxes is connected to the PC via the second RS232 serial port of the PC.

The glove only measures spreading of fingers, not the absolute sideways finger movement or position. This should not be a problem when analyzing gestures. The wrist joint has a large bend radius, so the wrist flex and wrist abduct sensors need relatively careful calibration. Most other sensors measure the bending of joints with a small bend radius. As long as the sensor is long compared to the bend area, little calibration is needed. Finally, the thumb movement is so complex that it is hard to capture in terms of few bend angles. Even with good calibration, the relative position of the thumb with respect to the fingers will not fit the real hand shape very well. This has to be taken into account for sign language analysis.

The glove electronics use analogue low-pass filters to attenuate the spectral components above 30 Hz at 20 db per decade. The documentation tells that human finger motion has been found to be usually slower than 3 Hz (that is, three movement cycles per second). The default sampling rate is 90 Hz, which is close to the maximum. The limiting factors are the analogue to digital conversion time and the serial data transfer. The high sampling rate is assumed to be useful to pinpoint the onset time of a movement. Sensor values will be integers in the range of 40 to 220 for most hands. For most sensors this means a resolution of about two units per degree.

The tracker electronics have a default setting of 104 Hz sampling rate. Higher rates (up to 144 Hz) are possible but not recommended. Both location and orientation are measured. Measures are represented as signed 14 bit integers which, at the default 91 cm range coordinate system, gives a resolution of roughly 1 mm. Angular resolution is about 0.1 degrees. The accuracy is about 0.5 degrees / 2.5 mm RMS averaged over the translational range.

3.2. Software

The trackers come with a text based configuration tool called CBIRD. For the glove and for the kinematics model of the hand, we use the Immersion Virtual Hand toolkit. As the Virtual Hand software does not support multiple tracker setups well, the CBIRD software has to be used to configure the Flock of Birds hardware manually before the Virtual Hand software can use both trackers simultaneously. The Immersion software consists of a Device Manager that manages hardware connections, the DCU configuration tool and a library (which comes with some sample programs).

Defaults are not well suited for use with two trackers, so both CBIRD and DCU have to be used to get the proper configuration after booting the hardware. In addition, the communication protocols are well documented. The advantage in using the Virtual Hand toolkit is that it contains a kinematic model of the hand geometry and movement abilities and a (mostly undocumented) interface to that.

The GloveGUI software simultaneously logs and visualizes the data from the trackers and the glove. Several aspects of the logging (like precision and speed) can be configured through command line options. The effective data rate is limited by the polling speed of Device Manager, as GloveGUI fetches the measurements from Device Manager. The visualization in turn fetches coordinates from the logged data. Default rates are 20 window updates and 25 log samples per second, but one will usually use higher log rates for sign language analysis.

The graphical user interface of GloveGUI is designed to minimize interaction: The user has to decide about log file processing (overwrite, append, abort) and has to decide whether to proceed if not all data sources are ready. Normally, the user only has to confirm the current setup to start the recording. To end the recording, it is enough to close the main window or hit the ESC key.

During recording, the three mouse buttons can be used to toggle the usage of tracker location information, to centre the hand display, and to cycle through the available viewing directions.

The log files are plain text, and the first sample of each log file is annotated with comment lines. Each sample consists of the following data: PC timestamp, glove electronics timestamp, switch/LED state, the 22 glove sensor bend angles (3 for each finger, 2 for the wrist, further sensors for the spread between fingers etc), tracker data (3D location, 3D direction). The 3D directions are available as a vector and as a triple of azimuth, elevation and roll. The following derived values are available: wrist location and rotation, end points of each finger bone (20 coordinate triples), and the movement distance for each point compared to the previous sample. The finger bone end point logging can be disabled with a command line option of GloveGUI.

4. The extension of ELAN to cover the analysis of numeric data

4.1. The ELAN annotation tool

ELAN is a linguistic annotation tool for digital video and or audio, which is available for several operating systems: Linux, Windows, and MacOS X. It provides integrated video/audio playback and supports the association of up to four video files with the annotation document. An unlimited number of annotation layers or tiers can be created. Typically, a tier groups annotations that describe the same phenomenon and that share the same constraints on structure and or content.

Two main concepts in ELAN's design are "media players" and "viewers".

ELAN creates a media player for each associated audio/video file. If media files are out of sync, they can be synchronized to each other within ELAN. ELAN provides a rich set of player controls to navigate through the media. The smallest step unit is 1 millisecond, which means that the media play head can be positioned with a maximum precision of 1 ms. Consequently, the boundaries of an annotation can be determined with millisecond precision.

The sample frequency of video is in most cases 25 or 30 frames per seconds, resulting in a frame duration of 40 or 33 ms. For these type of media, ELAN enables the user to step through the media frame by frame. Here the millisecond precision would not add anything. The higher frequency of the kinematic recordings described in section 3 better exploits ELAN's precision. As was described in the preceding section, kinematic recordings

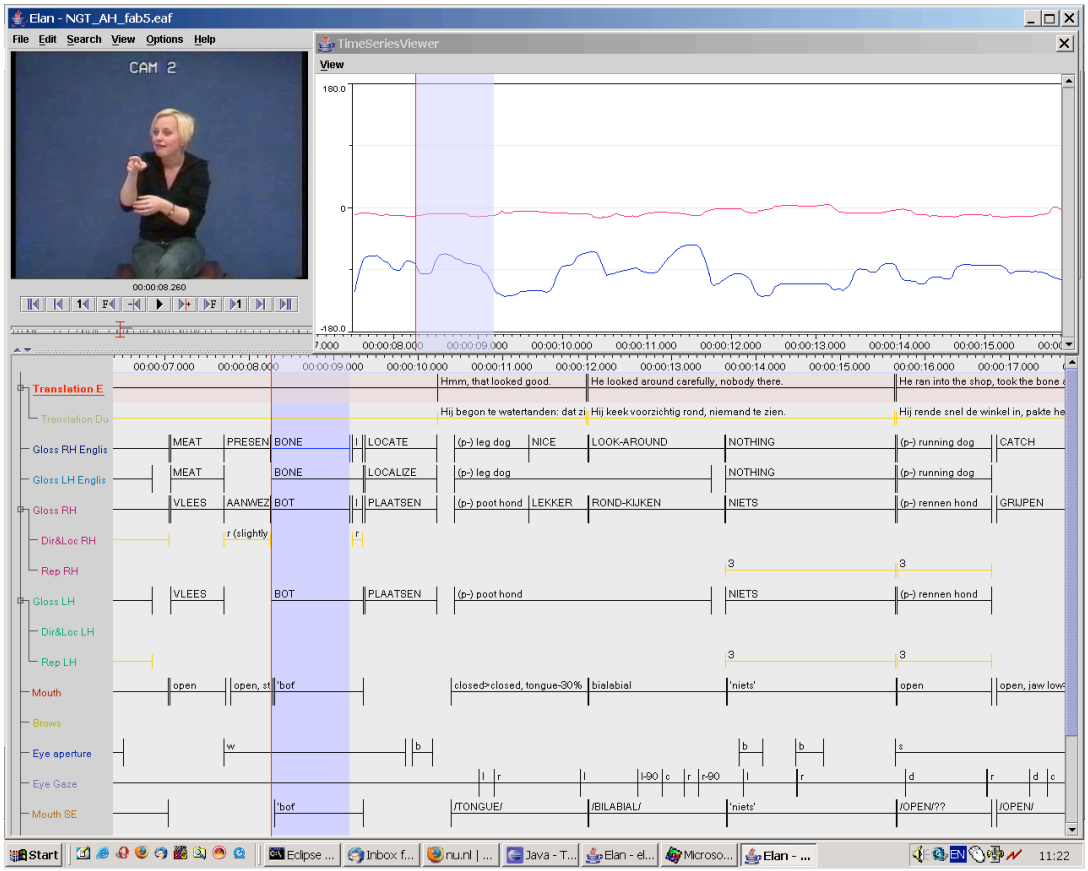

Figure 1: ELAN's media (top left), time series (top right), and timeline (bottom) viewers.

typically contain data with a frequency of around 100 Hz. This implies a significant increase in temporal precision for researchers studying sign languages. At this moment, ELAN offers no controls to navigate through the media in steps of 5 or 10 milliseconds: the user has to use the millisecond controls.

ELAN can display several viewers (see Figure 1), most of which present the underlying annotation data in their own specific way. However, there is also a viewer that can display a waveform of an audio file. All viewers are synchronized, i.e. they are connected to the media time and they all have some way of indicating the current media time and the selected time interval.

4.2. Integration of CyberGlove data in ELAN

A CyberGlove data file is associated with an annotation document in the same way as an audio or video file. Instead of a player, a specialized viewer is created to visualize the content extracted from the CyberGlove samples.

This viewer can contain multiple panes, and each pane can contain multiple movement tracks; the user is able to configure these.

4.2.1. A specialized reader for the proprietary CyberGlove file format

The CyberGlove software produces samples that consist of 40 - 100 distinguishable measurement values (see the description in section 3). A specialized reader or import module that is aware of the structure of such file has been developed, which is capable of dealing with the variations that can occur in this kind of files and is able to calculate the sample frequency from the timestamps in the samples: the sample frequency is not explicitly listed in the data file. For each field in the samples, the reader is able to create a track. A track consists of an array of (single) measurement values and some visualization attributes.

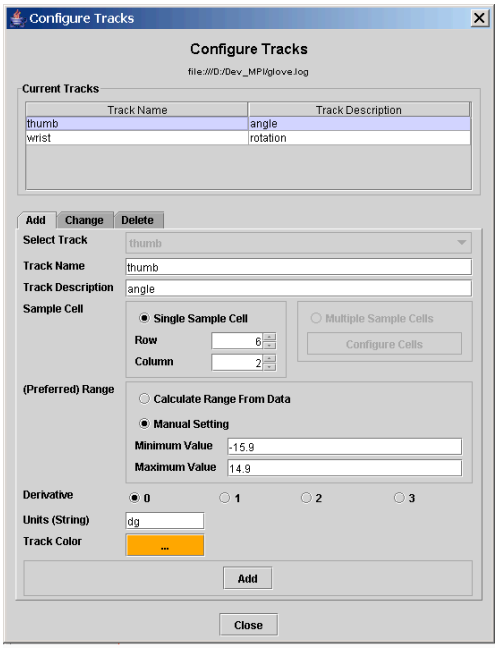

Figure 2: The track configuration pane.

4.2.2. Facility for track selection and calculation of derivatives

Given the multitude of measurement fields per sample, it is not feasible to simply visualize all information captured in the samples. Users need the opportunity to compose their own selection of tracks, based on the particular interest at hand. Therefore, a user interface has been developed to enable the selection and customization of tracks (Figures 2 and 3). Often the interest of the researcher goes beyond the bare data available in the file. For example, the amount of change over time in a certain measurement field could be equally important as the measurement itself, leading to the need for derived tracks. From a track recording the movement (covered distance) of a certain point on the glove, tracks for velocity, acceleration and jerk can be derived. For jerk, an extra filtering step might be necessary to reduce the noise in this type of derived data.

4.2.3. Facility for synchronization of the glove data to the media file(s).

Since it is very unlikely that the video files and CyberGlove recordings start at exactly the same time, the two signals need to be synchronized in some way. For that reason, the synchronization facility that already had been implemented in ELAN for audio and video files has been extended to support synchronization of video and CyberGlove streams. Corresponding events in both streams can be identified based on the graphical representation of the CyberGlove data. Thus, a new time origin for one or both streams can be determined and stored, guaranteeing synchronous playback in ELAN. To facilitate this process, use can be made of the button on the CyberGlove that switches a LED on and off: this will be visible in the video recordings, and changes the value of one of the parameters in the log file from 0 to 1 or the other way round (which can be visualised in one of the tracks).

4.2.4. Synchronized/connected viewer for CyberGlove data (time series viewer)

The tracks extracted from the CyberGlove file can be visualized as line plots, parallel to a (horizontal) time axis. A new viewer had to be created for this kind of data, a Timeseries viewer. The Timeseries viewer has a

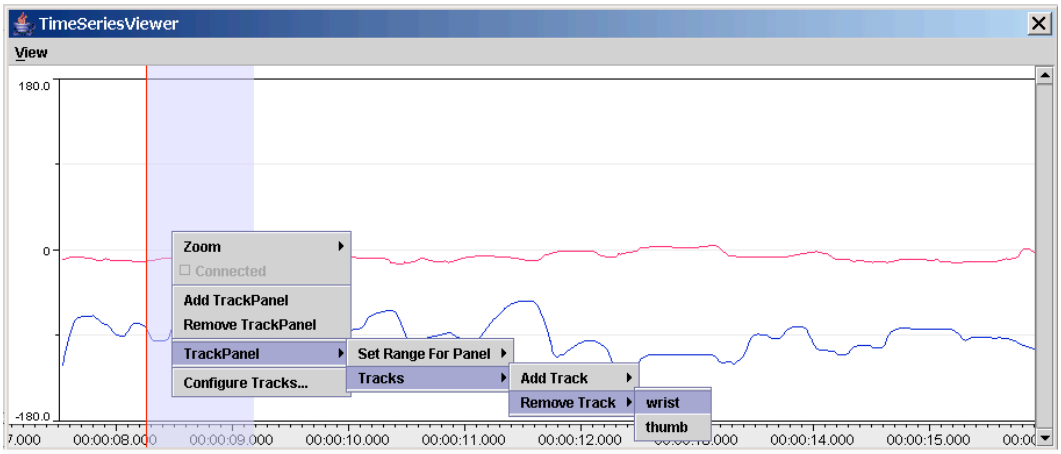

Figure 3: A pop-up menu to adjust what is displayed in the time series viewer.

horizontal time ruler, like the ELAN Timeline- and Waveform Viewer. The horizontal resolution can be changed by zooming in or out, effectively changing the number of milliseconds each pixel on the screen represents.

In addition to a horizontal ruler the time series viewer has also a, simple, vertical ruler labeling the minimum (low value) and maximum (high value) of the viewer area and a few values in between. A requirement specific to this viewer comes from the variety of value ranges that tracks can have. A track holding rotation values in degrees might have a range from -180 to 180 (or from 0 to 360) while a track representing velocity values might have a range from 0 to 20 and yet another track might have values between 0 and 1. The usability of the visual representation of the tracks would vanish when all these tracks would have to share the same coordinate system. To avoid this, the viewer can have multiple areas or 'panels', each one of which can display multiple tracks that can reasonably share the same range of amplitudes (coordinate system). Tracks can be added to and removed from a panel, panels can be added to or removed from the viewer (see Figure 2).

All viewers in ELAN have some way to visually mark the selected interval and the crosshair position (current media time). The Timeseries viewer marks these entities in the same way as other viewers with a time ruler, the Timeline viewer and the Waveform viewer (a light blue rectangle for the selected interval and a vertical red line for the crosshair).

In the Timeseries viewer the selected interval can be used for the extraction of the minimum or maximum value within that interval on a certain track.

4.3. Integration of glove data with annotations

The benefit of the integrated visualization of numeric data with video and annotations is twofold: on the one hand the line plots can be assistive in accurate annotation creation, on the other hand numeric values can be transferred to annotation values. The latter feature is very important for quantitative research.

This transfer has been implemented as follows: for each annotation on tier X, extract minimum or maximum in the annotation's interval from track Y and add the result to a (new) child annotation on tier Z (being a child tier of tier X). The resulting annotations can be included in the export to tab-delimited text for further processing in a spreadsheet or in an application for statistical analysis.

5. Future developments and use

5.1. Inclusion of other data formats

The initial efforts in the area of integration of numeric data in ELAN have been geared towards full support for the very specific data files generated by the *Cyberglove* data glove and *Flock of Birds* motion sensors. However, in the design and implementation stage, a much broader application has always been in mind.

The components developed for the CyberGlove data are suited for any kind of 'time series' data, i.e. any data consisting of a list or array of time - value pairs, produced by whichever device or software application. Eye tracking equipment is just one example.

The main problem here is that there does not seem to be a standard format for such data. It is therefore unavoidable to write special 'readers' or import modules for each kind of file. Such modules have been created for the CyberGlove files and for a very simple, generic time - value file, a text file where each line holds a timestamp and a value, separated by a white space.

A Service Provider interface has been defined to enable addition of other, custom import modules. Third parties should be able

to create and add their own module to the software independently from ELAN release dates.

5.2. Addition of phonetic data from spoken languages

ELAN's new functionality extension to represent and present time series data can also be seen as an opening to include phonetic data from spoken languages. For a number of research programs it becomes increasingly interesting to combine different data streams for analysis. In phonetics, it is interesting to combine the speech wave with video recordings of lip movements and other visible articulators, and with laryngograph and articulograph recordings, for example. ELAN makes it possible to easily synchronise these streams and present them in a synchronous presentation together with layers of annotation. In this respect ELAN offers unique functionality.

5.3. Linguistic uses

For the type of phonetic research on sign languages characterised in section 2, the present enhancement of ELAN offers two separate types of benefits. Firstly, at a practical level, the visualisation of the articulatory (kinematic) recordings in parallel with the video recordings allows for very efficient segmentation of these kinematic data: the test items can be quickly identified in the video window, and then precisely segmented using the higher temporal resolution of the movement tracks.

Secondly, in addition to the targeted recording and analysis of experimental phonetic data, a quite different type of use is also envisaged. Linguists working on phonological and prosodic aspects of sign languages can use the visualisation of the arm, hand and finger movements to generate research questions and hypotheses about various aspects of the form of signing. It has been difficult to develop sensible notions about movement features such as size and 'manner' (tenseness, speed), as well as 'holds' (pauses), without a common view on what exactly these properties refer to. While the phonological categories need not necessarily be expressed in terms of kinematic features, the ability to explore sets of video data with a view on displacement, velocity and acceleration of selected points on the hand is foreseen to have great benefits and lead to new insights on the form of connected signing.

In this respect, a great advantage of the present approach to accessing and processing data from movement trackers is that the data are stored and integrated in a multimedia corpus, rather than being data that are only used for an experiment and then disregarded. Metadata that characterise the overall properties of the recording situation (as in the IMDI standard, for example; see Broeder & Offenga 2004) are flexible enough to include a description of the general properties of the kinematic recordings, including a reference to the data file. In this way, the wider use of the data beyond the experiment at hand is indeed a feasible option.

6. References

Brentari, D. (1998) *A prosodic model of sign language phonology*. Cambridge, MA: MIT Press.

Broeder, D. & F. Offenga (2004) IMDI metadata set 3.0. *Language Archive Newsletter*, 1-2: 3.

Brugman, H., O. Crasborn & A. Russell (2004) Collaborative annotation of sign language data with peer-to-peer technology. In: *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation* (pp. 213-216). M.T. Lino et al., eds.

Cheek, A. (2001) *The phonetics and phonology of handshape in American Sign Language*. PhD thesis, UT Austin.

Crasborn, O. (2001) *Phonetic implementation of phonological categories in Sign Language of the Netherlands*. Utrecht: LOT. (PhD thesis, Leiden University.)

Crasborn, O. (2006) Nonmanual structures in sign languages. In: *Encyclopedia of Language and Linguistics, 2nd ed.* (pp. 668-672) K. Brown, ed. Oxford: Elsevier.

Crasborn, O., E. van der Kooij, D. Broeder & H. Brugman (2004) Sharing sign language corpora online: proposals for transcription and metadata categories. In: *Proceedings of the LREC 2004 Satellite Workshop on Representation and processing of sign languages* (pp. 20-23) O. Streiter & C. Vettori, eds.

Stokoe, W. (1960) *Sign language structure. An outline of the visual communication systems of the American Deaf.* Silver Spring, MD: Linstok Press.

Wilbur, R. (1990) An experimental investigation of stressed sig production. *International Journal of Sign Linguistics* 1: 41-59.

Wilbur, R. & S.B. Nolen (1986) The duration of syllables in American Sign Language. *Language and Speech* 29: 263-280.

Wilcox, S. (1992) *The phonetics of fingerspelling*. Amsterdam: John Benjamins.