Taking non-manuality into account in collecting and analyzing Finnish Sign Language video data

Anna Puupponen, Tommi Jantunen, Ritva Takkinen, Tuija Wainio, Outi Pippuri
Department of Languages (Sign Language Centre), University of Jyväskylä, Finland
P.O. Box 35, FI-40014 University of Jyväskylä, Finland
E-mail: {anna.puupponen, tommi.j.jantunen, ritva.a.takkinen, tuija.wainio}@jyu.fi & outi.k.pippuri@student.jyu.fi

Abstract

This paper describes our attention to research into non-manuals when collecting a large body of video data in Finnish Sign Language (FinSL). We will first of all give an overview of the data-collecting process and of the choices that we made in order for the data to be usable in research into non-manual activity (e.g. camera arrangement, video compression, and Kinect technology). Secondly, the paper will outline our plans for the analysis of the non-manual features of this data. We discuss the technological methods we plan to use in our investigation of non-manual features (i.e. computer-vision based methods) and give examples of the type of results that this kind of approach can provide us with.

Keywords: Finnish Sign Language, non-manuals, video data, Kinect, SLMotion, head movements

1. Introduction

This paper describes the process of collecting high quality video data in Finnish Sign Language (FinSL) and how, in the process, we took into account the investigation of non-manual elements (i.e. the movements and positions of the head and torso, eyes, eye brows, and mouth). The paper also describes how we plan to analyze the non-manual elements in the data. We present a technological method that has been specifically developed for such an analysis and, in addition, demonstrate how this method has already been used in the phonetic and linguistic analyses of FinSL head movements.

The data collection and the work with non-manuals are directly motivated by two research projects presently being carried out in the Sign Language Centre of the University of Jyväskylä, Finland. The first is the *FinSLs Corpus* project, which aims to build a high quality video corpus for the sign languages of Finland[1]. The second is the *ProGram* project, which aims to investigate the syntax and prosody of FinSL[2]. Both projects are closely linked to other current Finnish projects dealing with data collection and technological methods, most notably the *Corpus and Sign Wiki* project[3] and the *CoBaSiL* project.[4]

Large video corpora on sign languages have traditionally been collected and analyzed only in terms of manual activity (e.g. Crasborn & Zwitserlood 2008b; Johnston 2009; Wallin et al. 2010). No widely used standards for collecting and analyzing non-manual elements exist and, consequently, when non-manual elements have been investigated systematically, researchers have had to investigate them on the basis of video material that was not specifically recorded for the purpose. For the specific analysis of non-manual elements, technological methods have long depended on various utilizations of motion capture technology (e.g. Jantunen et al. 2012; Puupponen et al. 2013). However, recent developments in computer vision and image analysis techniques have also made it possible to deploy content-based video analysis methods for research into non-manuals (e.g. Karppa et al. 2011; Luzardo et al. 2013).

In the rest of this paper, Section 2 presents the collecting and processing of high definition (HD) video material, with particular emphasis on research into non-manuals in sign languages. Section 3 describes how the data can and has been used in the analysis of non-manuals. Section 4 offers a brief conclusion.

2. Video data on FinSL

2.1 Background

At the Sign Language Centre in the University of Jyväskylä, we aim to collect a corpus of FinSL and Finland-Swedish Sign Language (FinSSL). Currently, our data consists of 10 hours of multi-camera HD (1920x1080) 25-50 fps video material on FinSL, recorded in the Audio-visual Research Centre at the University of Jyväskylä. We have collected material from a total of seven pairs of informants (age 20 to 59 years) from different parts of Finland, who all performed a fixed series of seven tasks. The data includes both dialogue and monologue material.

The procedure for data collection mainly follows the conventions of earlier corpus projects of several other sign languages, e.g. German Sign Language (Hanke et al. 2010), the Sign Language of the Netherlands (Crasborn & Zwitserlood 2008a), and Swedish Sign Language (Mesch 2009). In the procedure, two signers take part in a conversation in which they first talk about themselves, their work, their hobbies or something they are interested in. The signers then take it in turns to sign from comics and tell a story from a picture book, and finally they discuss an issue that concerns the Deaf world or FinSL.

[1] http://viittomakielenkeskus.jyu.fi/projektit.html
[2] http://users.jyu.fi/~tojantun/ProGram/index.html
[3] http://www.kl-deaf.fi/fi-FI/Korpus-SignWiki/
[4] http://research.ics.aalto.fi/cbir/cobasil/

The material will be annotated in ELAN (Crasborn & Sloetjes 2008). At the time of writing, the annotation of manual activity has just begun and the annotation of non-manuals is scheduled to begin in the autumn of 2014. Also metadata is being gathered and indexed according to the IMDI standard. The metadata consists of age, sex, place of residence, school, education, age of sign language learning, languages used at home, work, languages used at work etc.

The video material, annotations and metadata are stored in Jyväskylä University's quota in the IDA storage service provided by the CSC – IT Center for Science[5]. The service will make it possible, for example, to publish the data for the use of the general public in the future.

2.2 Recording and processing the data

For the current data, the camera set-up consisted of seven Panasonic HD video cameras, illustrated in Figure 1. Of the cameras, six (Cams. 1-6) were directed towards the informants and one recorded the person giving the instructions (Cam. 7). Cam. 1 recorded an image of both of the informants (Signers A & B) facing each other, whereas Cam. 2 and Cam. 3 recorded an image of each of the signers from approximately a 45-degree angle.

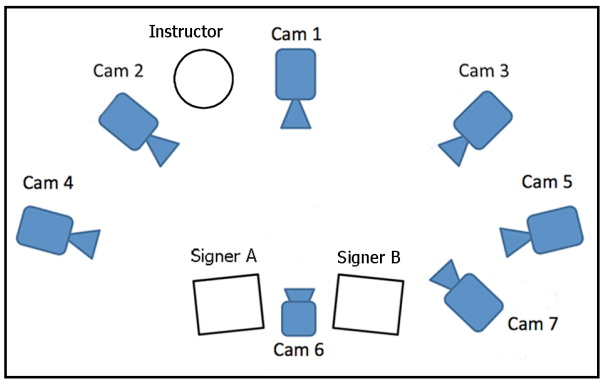

Figure 1: Camera arrangement in the recording of FinSLs corpus material.

In order to collect high quality material for research into non-manual activity, i.e. to observe more closely the movements and positions of the torso, head, and face, we had extra cameras recording close-up views of the upper body of both informants (Cams. 4 and 5). For these close-ups the cameras were positioned directly in front of the informants, so that they recorded a nearly direct image of the signers. The direct image footage with a front view of the signers makes possible computer-vision based analysis of non-manuals (see Section 3 of this paper).

To aid the analysis of the informants' signing in the dimension of depth, we had one camera recording the informants from above (Cam. 6). This camera was attached to the ceiling of the studio. Example frames of the video material from four different camera angles are presented in Figure 2.

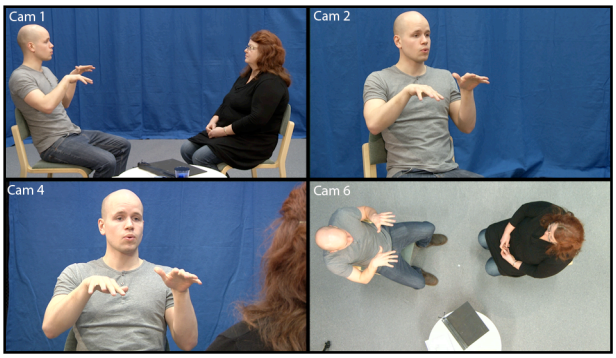

Figure 2: Screenshots of one frame in the video material from the recordings of Cameras 1, 2, 4, and 6.

The recorded video material was subsequently edited in *Adobe Premiere Pro CS6 6.0.5*. To help the editing process of multi-angle video clips, a clapperboard was used at the beginning and end of each of the seven tasks in all of the dialogues. The video material was edited so that the footage from six different camera angles (Cams. 1–6) resulted in six separate synchronised video clips for each task. The recordings of the person giving the instructions (Cam. 7) did not contain the clapperboard signals, and was therefore edited and compressed separately into one continuous clip, containing all the instructions for the different tasks in each dialogue.

All the recordings have been stored in Material eXhange Format (MXF) and compressed using H.264 in an MP4 container. The MXF container format contains time-code and metadata support and is not specific to a compression scheme. It is being used for the storage of the material to avoid restrictions in future compression. H.264 compression was used to ensure usability and compatibility between different operating systems when annotating the material in ELAN. The video material was compressed so that the annotation and analysis of both manual and non-manual activity could be done on the basis of HD material and with a reasonable file size.

2.2 Additional Kinect data

In addition to the HD video, our current material also includes data recorded with one *Kinect* motion sensing input device. In the studio, the device was stationed next to Cam. 2, where it always recorded the activity of one of each pair of informants (Signer B in Figure 1). The device was connected to an Apple MacBook Pro 15" laptop (2,6 GHz Intel Core i7) and controlled with specifically coded *NiRecorder* software, based on *OpenNi*[6] and *SensorKinect*[7] technologies. All the recordings have been stored on the hard drive of the laptop.

[5] http://www.csc.fi/english

[6] http://www.openni.org

[7] http://github.com/avin2/SensorKinect

The purpose of recording *Kinect* data was to complement the main HD video data, especially with quantitative information about depth, a dimension not inherently present in traditional video recordings. In practice, the *Kinect* data consist of a low-quality RGB video, augmented with a 16 Hz infrared video, and an automatically calculated skeleton model of the signer. Of these, the infrared video, shown in Figure 3, allows one to investigate the activity of signers in the dimension of depth to the precision of one millimetre. From the point of view of non-manuals, such data will be particularly useful in the analysis of the depth of head and body movements and postures which will be carried out in the *ProGram* project. More generally, when combined with data recorded with Cam. 6 the data make possible a very precise analysis of, for example, the spatial relationship of the hand and the rest of the body.

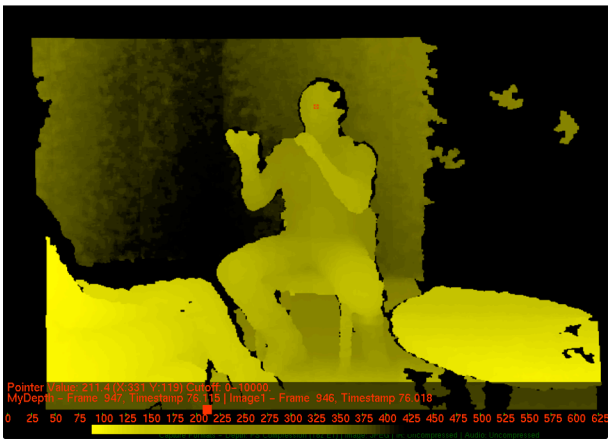

Figure 3: A screenshot showing a frame from the infrared data. The pointer value indicates the distance of the signer's forehead from the *Kinect* sensor.

The skeleton data adds further value to the analysis of the signers' movements as it provides data analogous to that collected with traditional motion capture (mocap) equipment (Chen & Koskela 2013). The skeleton figure, illustrated in Figure 4, is extracted on the basis of the depth data in real time during the recording. In practice, the skeleton figure gives a three-dimensional model of the global movements of the arms, legs, torso, and head of the signer.

The extraction of the skeleton figure is based on an algorithm that classifies a large three-dimensional point cloud into approximately a dozen human skeleton joint coordinates (Chen & Koskela 2013). This data is stored as a Comma Separated Value (CSV) file that can be easily imported to common mathematical software, such as *Matlab*, for further analysis. In terms of non-manuality, the skeleton data allows one to analyze such matters as the kinematic properties of global movements of the head and torso with a methodology developed for mocap studies (see Jantunen et al. 2012). Again, such work is planned to be carried out in the *ProGram* project.

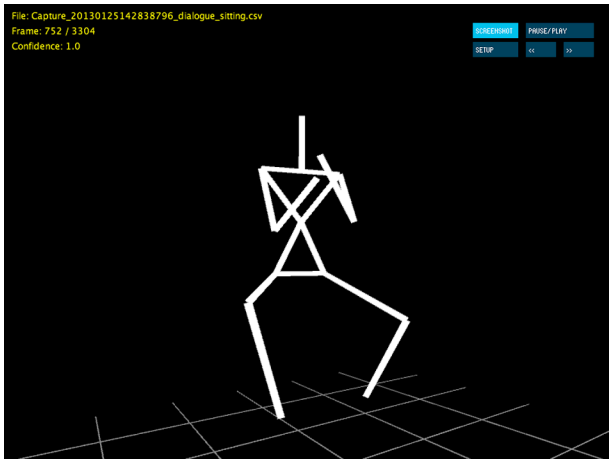

Figure 4: A three-dimensional skeleton model of a signer. The model can be viewed from different angles and from different distances.

3. Analyzing the data

3.1 SLMotion software

The high quality multi-camera video data will allow us to investigate non-manuality not only using traditional observation methods but also with various computer-vision based image analysis technologies. The main technologies that we are going to use are included in the *SLMotion* software (Karppa et al. 2014).[8] SLMotion is a tool for a near mocap-quality motion analysis of various articulators of signers visible in videos containing sign language. The first development versions of the tool focused on the hands and the head, and measured the motion of these articulators by first detecting parts of the person's bare skin on a video, then characterizing the shapes of the hands and the head with a point distribution model, and finally tracking their motion separately by the Kanade-Lucas-Tomasi algorithm and active shape models (Karppa et al. 2011). Recent development work has added to the tool the functionality to track and measure the movements and positions of the eyes, eye brows, and mouth (Luzardo et al 2013, 2014). A useful feature of SLMotion is that the quantitative results produced with it can be imported into ELAN for visualization and further analysis. This is illustrated in Figure 5 for the present data.

All the basic functions of SLMotion will be utilized in the *ProGram* project from 2015 onwards. Concerning non-manual activity, the tool will be used both as an aid to annotation and for the quantitative analyses of movements produced by the torso and the head. With respect to annotation, the ability of the tool to detect and classify, for example, eye blinks can be used to automate the manually time-consuming annotation process. Concerning the activity of the torso and head, the project will

[8] http://users.ics.aalto.fi/jmkarppa/slmotion/

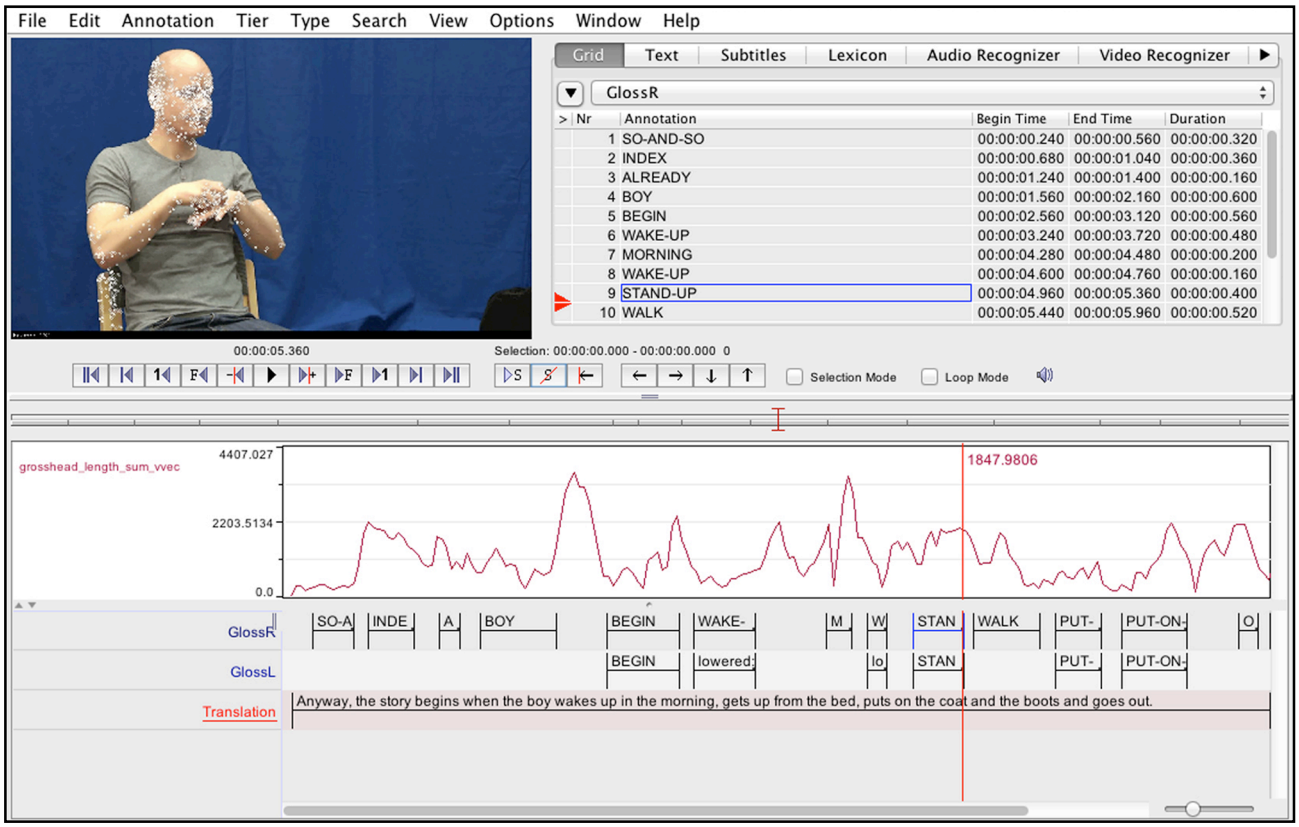

Figure 5: A screenshot from ELAN showing visualized SLMotion speed data of the movement of the head

focus on investigation of the signers' sentence-internal body and head movements and analyse them with various correlation functions for rhythm (see Jantunen et al. 2012). The quantitative data for this will be provided by the latest functions of SLMotion (see Luzardo et al. 2013).

Although non-manual elements have not yet been systematically investigated in the present data, we have already used the earlier development stages of SLMotion in the analysis of non-manual activity in FinSL (e.g. Jantunen et al. 2010; Puupponen 2012). In the following, we give examples of the results that this kind of technologically oriented analysis of sign language can produce. In particular, we will focus on head movements in horizontal and vertical dimensions.

3.2 On the head movements in FinSL

In our study of horizontal and vertical head movements in FinSL, we used traditional observation methods and SLMotion-based analysis to examine the phonetic forms and linguistic functions of articulations produced with the head (Puupponen 2012). The visualized SLMotion measurement data was found useful, for example, in identifying a particular head movement (e.g. a headshake) from the continuous stream of head movements, as well as in defining the starting and ending point of different head movements. Also a more detailed segmentation of head movements and an investigation of differences between head movements of a certain type were carried out on the basis of the numerical data.

In the data discussed in Puupponen (2012), seven different types of head movement were identified. Of these, five were included in our analysis: *nod*, *nodding* (a series of small repeated nods), *head turn*, *sideways tilt of the head*, and a *headshake*. Nodding movements and head shakes were repeated movements consisting of six to seven movement phases, whereas head nods, head turns and sideways tilts were non-repeated movements consisting of one to three movement phases. The two excluded movement types, *head thrust* and *backward pull of the head*, were produced in the dimension of depth. Because of the two-dimensionality of the video, the phonetic description and analysis of these movements was not possible with SLMotion at that time.

In general, the analysis showed that the head is very active during signing, as is demonstrated in Figure 6. It was argued in Puupponen (2012) that the continuous movement produced by the head has consequences for the annotation and analysis of head movements: identifying the head movements and distinguishing, for example, between linguistic and non-linguistic elements from the continuous stream of head movements is not always clear-cut (see also Puupponen et al. 2013).

Even though both articulators were moving continuously, the SLMotion measurements for the horizontal and vertical movements of the head did not correlate to those of the dominant hand ($r \leq 0{,}3$ in all cases). In addition, the head movements in the data were often not temporally aligned with the manually produced signing sequences (i.e. syntactic constituents).

Concerning the movement-internal features of different head movements, SLMotion analysis revealed that in most of the head movements involving repetition the amount of movement increased at the start and then decreased towards the end. This motion diminution is a feature associated with both horizontal headshakes and vertical nodding movements. The phenomenon is demonstrated in Figure 7.

The different types of head movements in the data signalled, for example, assertion and affirmation (nod, nodding), negation, semantic exclusion and hesitation (head turn, headshake, head tilt), and the end of a topic phrase (nod). The head movements also appeared at the beginning of text episodes (nod, nodding) and they made the signing textually and syntactically coherent by making meaningful use of the three-dimensional signing space (head tilt). Also in many cases two head movements occurred simultaneously. This was particularly a quality of head tilts, a fact possibly resulting from the long duration and textual-syntactic functions of head tilts, which allow the production of simultaneous head movements with, for example, emphatic functions.

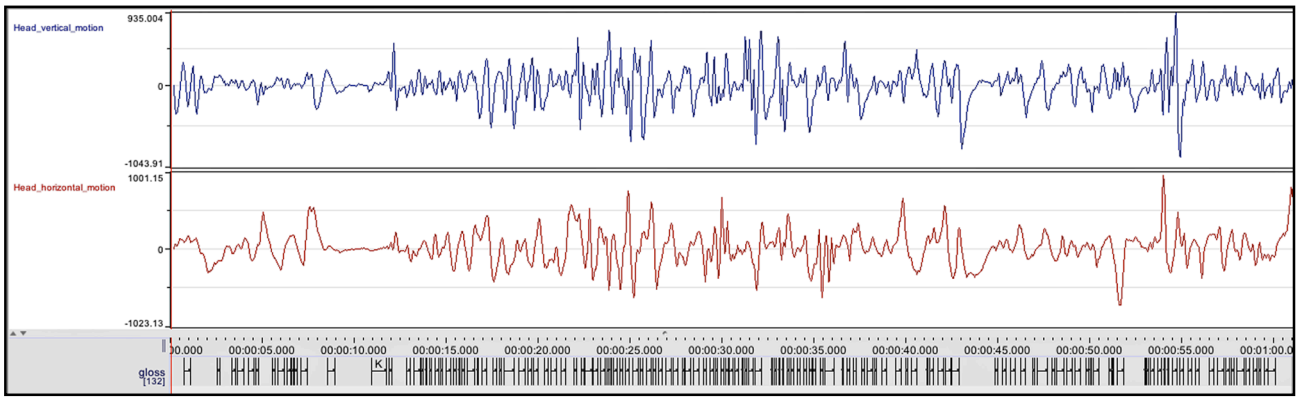

Figure 6: A screenshot from ELAN showing overall visualisations of the horizontal and vertical motion of the head in the data. The annotations of manual signs are shown in the annotation tier below the graphs.

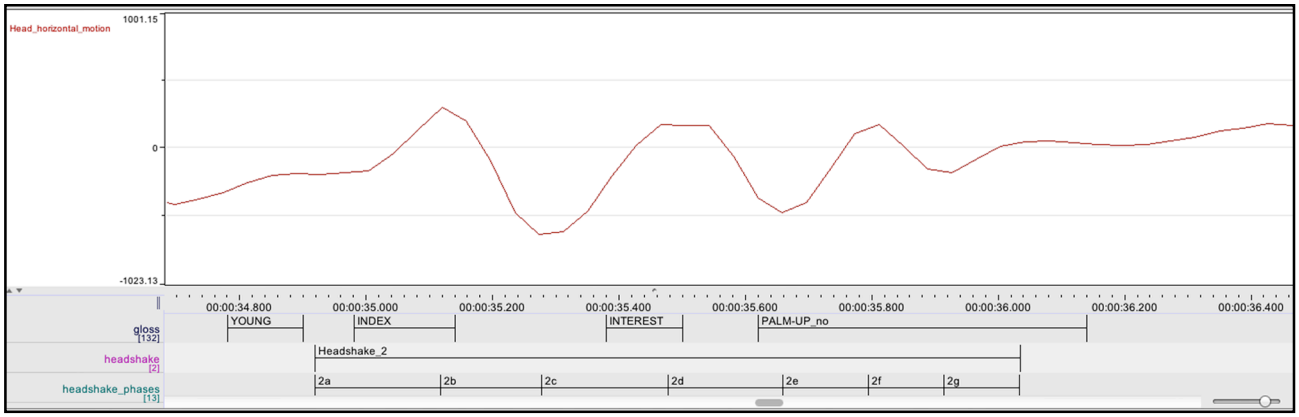

Figure 7: A screenshot from ELAN showing the visualised data of the horizontal motion of one headshake. The annotations for the manual signs, head movements, and movement-internal phases are shown on tiers below the graph.

4. Conclusion

In this paper we have discussed how we collected multi-camera HD video and Kinect material in FinSL, with particular reference to non-manuals. We have also introduced our already existing analyses of non-manuals in this type of data and presented our plans for the future. Although the analysis of non-manuals in the data that we have recently gathered for that purpose has not yet properly begun, we are convinced that knowledge of the processes we have described in this paper will also be of benefit to others working in the field.

5. Acknowledgements

The authors wish to thank Eleanor Underwood for checking the English of the paper. The financial support of the Academy of Finland under grants 269089 and 273408 is gratefully acknowledged.

References

Chen, X. & Koskela, M. (2013). Online RGB-D gesture recognition with extreme learning machines. In the *Proceedings of the 15th ACM International Conference on Multimodal Interaction* (ICMI'13), Springer Verlag, pp. 467-474.

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In the *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora* [organized as a part of LREC 2008, Paris: ELRA, pp. 39-43.

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

Crasborn, O. & Zwitserlood, I. (2008b). *Annotation of the video data in the "Corpus NGT"*. Dept. of Linguistics & Centre for Language Studies, Radboud University Nijmegen, The Netherlands. Online publication http://hdl.handle.net/1839/00-0000-0000-000A-3F63-4 (accessed 15 January 2013).

Hanke, T., König, L., Wagner, S. & Matthes, S. (2010). DGS Corpus & Dicta-Sign: The Hamburg Studio Setup. In the *Proceedings of the 4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Paris: ELRA, pp. 106-109.

Jantunen, T., Koskela, M., Laaksonen, J. & Rainò, P. (2010). Towards the automated visualization and analysis of signed language motion: method and linguistic issues. In the *Proceedings of the 5th International Conference on Speech Prosody* (SP 2010), 100006:1-4.

Jantunen, T., Burger, B., De Weerdt, D., Seilola, I. & Wainio, T. (2012). Experiences collecting motion capture data on continuous signing. In the *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon*, Paris: ELRA pp. 75-82.

Johnston, T. (2009). *Guidelines for annotation of the video data in the Auslan Corpus*. Dept. of Linguistics, Macquarie University, Sydney, Australia. Online publication http://media.auslan.org.au/media/upload/attachments/Annotation_Guidelines_Auslan_CorpusT5.pdf (accessed 15 January 2013).

Karppa, M., Jantunen, T., Koskela, M., Laaksonen, J. & Viitaniemi, V. (2011). Method for visualisation and analysis of hand and head movements in sign language video. In the *Proceedings of the 2nd Gesture and Speech in Interaction conference* (GESPIN 2011). [CD]

Karppa, M., Viitaniemi, V., Luzardo, M., Laaksonen, J. & Jantunen, T. (2014). SLMotion: An extensible sign language oriented video analysis tool. To appear in the *Proceedings of the 9th Language Resources and Evaluation Conference* (LREC 2014). Paris: ELRA.

Luzardo M; Karppa, M.; Laaksonen, J.; Jantunen, T. (2013). Head pose estimation for sign language video. In *Image Analysis Lecture Notes in Computer Science,* Vol. 7944, Springer Berlin Heidelberg, pp. 349-360.

Luzardo, M., Viitaniemi, V., Karppa, M., Laaksonen, J. & Jantunen, T. (2014). Estimating Head Pose and State of Facial Elements for Sign Language Video. To appear in the *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*. Paris: ELRA.

Mesch, J. (2009). Project Planning: the Swedish Sign Language Corpus. Presentation at the *Sign Linguistics Corpora Network Workshop 1: Introduction and Data Collection* in London, England, July 26-27, 2009.

Puupponen, A. (2012*). Horisontaaliset ja vertikaaliset päänliikkeet suomalaisessa viittomakielessä* [Horizontal and vertical head movements in FinSL], MA thesis, University of Jyväskylä, Jyväskylä, Finland. [http://urn.fi/URN:NBN:fi:jyu-201207242120]

Puupponen, A.; Jantunen, T.; Wainio, T. & Burger, B. (2013). Messing with the head: on the form and function of head movements in Finnish Sign Language. Presentation at the 11th Theoretical Issues in Sign Language Research conference (TISLR 11), University College London, July 10-13, 2013.

Wallin, L., Mesch, J. & Nilsson, A.-L. (2010). *Transcription guidelines for Swedish Sign Language discourse* (Version 1). Dept. of Linguistics, University of Stockholm, Sweden.