5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon

27 May 2012

ABSTRACTS

Editors:

**Onno Crasborn, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke,
Jette Kristoffersen, Johanna Mesch**

Workshop Programme

09:00 – 10:30	Oral Session A: *Relations between Corpus and Lexicon*
10:30 – 11:00	Coffee break
11:00 – 13:00	Poster Session B: *Methodology and Technology*
13:00 – 14:00	Lunch break
14:00 – 16:00	Poster Session C: *Resources*
16:00 – 16:30	Coffee break
16:30 – 19:00	Oral Session D: *Issues in the Construction of Sign Corpora*

Workshop Organizers

Onno Crasborn	Radboud University, Nijmegen NL
Eleni Efthimiou	Institute for Language and Speech Processing, Athens GR
Evita Fotinea	Institute for Language and Speech Processing, Athens GR
Thomas Hanke	Institute of German Sign Language, University of Hamburg, Hamburg DE
Jette Kristoffersen	Centre for Sign Language, University College Capital, Copenhagen DK
Johanna Mesch	Stockholm University, Stockholm SE

Workshop Programme Committee

Richard Bowden	University of Surrey, Guildford GB
Penny Boyes Braem	Center for Sign Language Research, Basel CH
Annelies Braffort	LIMSI/CNRS, Orsay FR
Christophe Collet	IRIT, University of Toulouse, Toulouse FR
Helen Cooper	University of Surrey, Guildford GB
Kearsy Cormier	Deafness Cognition and Language Research Centre, London GB
Onno Crasborn	Radboud University, Nijmegen NL
Eleni Efthimiou	Institute for Language and Speech Processing, Athens GR
Evita Fotinea	Institute for Language and Speech Processing, Athens GR
John Glauert	University of East Anglia, Norwich GB
Thomas Hanke	Institute of German Sign Language, University of Hamburg, Hamburg DE
Alexis Heloir	German Research Centre for Artificial Intelligence, Saarbrücken DE
Jens Heßmann	University of Applied Sciences Magdeburg-Stendal, Magdeburg DE
Matt Huenerfauth	City University of New York, New York US
Trevor Johnston	Macquarie University, Sydney AU
Reiner Konrad	Institute of German Sign Language, University of Hamburg, Hamburg DE
Jette Kristoffersen	Centre for Sign Language, University College Capital, Copenhagen DK
Lorraine Leeson	Trinity College, Dublin IE
Petros Maragos	National Technical University, Athens GR
Johanna Mesch	Stockholm University, Stockholm SE
Carol Neidle	Boston University, Boston US
Christian Rathmann	Institute of German Sign Language, University of Hamburg, Hamburg DE
Adam Schembri	National Institute for Deaf Studies and Sign Language, La Trobe University, Melbourne AU
Meike Vaupel	University of Applied Sciences Zwickau, Zwickau DE

Session A: Relations between Corpus and Lexicon

Sunday 27 May, 09:00 – 10:30
Chairperson: Onno Crasborn Oral Session

From Meaning to Signs and Back: Lexicography and the Swedish Sign Language Corpus

Johanna Mesch and Lars Wallin

In this paper, we will present the advantages of having a reference dictionary, and how having a corpus makes dictionary making easier and more effective. It also gives a new perspective on sign entries in the dictionary, for example, if a sign uses one or two hands, or which meaning 'genuine signs' have, and it helps find a model for categorization of polysynthetic signs that is not found in the dictionary. Categorizing glosses in the corpus work has compelled us to revisit the dictionary to add signs from the corpus that are not already in the dictionary and to improve sign entries already in the dictionary based on insights that have been gained while working on the corpus.

Linking an ID-gloss Database of ASL with Child Language Corpora

Julia Fanghella, Leah Geer, Jonathan Henner, Julie Hochgesang, Diane Lillo-Martin, Gaurav Mathur, Gene Mirus and Pedro Pascual-Villanueva

We describe an on-going project to develop a lexical database of American Sign Language (ASL) as a tool for annotating ASL corpora collected in the United States. Labs within our team complete locally chosen fields using their notation system of choice, and pick from globally available, agreed-upon fields, which are then merged into the global database. Here, we compare glosses in the database to annotations of spontaneous child data from the BiBiBi project (Chen Pichler et al., 2010). These comparisons validate our need to develop a digital link between the database and corpus. This link will help ensure that annotators use the appropriate ID-glosses and allow needed glosses to be readily detected (Johnston, 2011b; Hanke and Storz, 2008). An ID-gloss database is essential for consistent, systematic annotation of sign language corpora, as (Johnston, 2011b) has pointed out. Next steps in expanding and strengthening our database's connection to ASL corpora include (i) looking more carefully at the source of data (e.g. who is signing, language background, age, region, etc.), (ii) taking into account signing genre (e.g. presentation, informal conversation, child-directed etc.), and (iii) confronting the matter of deixis, gesture, depicting verbs and other constructions that depend on signing space.

Integrating Corpora and Dictionaries: Problems and Perspectives, with Particular Respect to the Treatment of Sign Language

Jette H. Kristoffersen and Thomas Troelsgård

In this paper, we will discuss different possibilities for integration of corpus data with dictionary data, mainly seen from a lexicographic point of view and in a sign language context. For about 25 years a text corpus has been considered a useful, if not necessary tool for editing dictionaries of written and spoken languages. Corpora are equally useful to sign language lexicographers, but sign language corpora have not become accessible until recent years. Nowadays corpora exist, or are being developed, for several sign languages, and we have the possibility of editing new, truly corpus-based sign language dictionaries, and of developing interfaces that integrate corpus and dictionary data. After a brief look at three existing integrated interfaces, one for German, one for Danish, and one for Danish Sign Language, we point out some of the problems that should be considered when making an integrated interface, and, finally, we briefly outline the future perspectives of integrated sign language corpus-dictionary interfaces.

Session B: Methodology and Technology

Sunday 27 May, 11:00 – 13:00

Chairperson: Johanna Mesch Poster Session

From Corpus to Lexical Database to Online Dictionary: Issues in Annotation of the BSL Corpus and the Development of BSL SignBank

Kearsy Cormier, Jordan Fenlon, Trevor Johnston, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam and Bencie Woll

One requirement of a sign language corpus is that it should be machine-readable, but only a systematic approach to annotation that involves lemmatisation of the sign language glosses can make this possible at the present time. Such lemmatisation involves grouping morphological and phonological variants together into a single lemma, so that all related variants of a sign can be identified and analysed as a single sign. This lemmatisation process is made more straightforward by the existence of a comprehensive lexical database, as in the case for Australian Sign Language (Auslan). When annotation of data collected as part of the British Sign Language (BSL) Corpus Project began, no such lexical database for BSL existed. Therefore, a lemmatised BSL lexical database was created concurrently during annotation of the BSL Corpus data. As part of ongoing work by the Deafness Cognition & Language Research Centre, this lexical database is being developed into an online BSL dictionary, BSL SignBank. This paper describes the adaptation of the Auslan lexical database into a BSL lexical database, and the current development of this lexical database into BSL SignBank.

Linking Corpus NGT Annotations to a Lexical Database Using Open Source Tools ELAN and LEXUS

Onno Crasborn, Micha Hulsbosch and Han Sloetjes

This paper describes how we have made a first start with expanding the functionality of the ELAN annotation tool to create a bridge to a lexical database. A first lookup facility of an annotation in a LEXUS database is created, which generates a user-configurable selection of fields from that database, to be displayed in ELAN. In addition, an extension of the (open) controlled vocabularies that can be specified for tiers allows for the creation of very large vocabularies, such as lexical items in a language. Such an ‘external controlled vocabulary’ is an XML file that can be published on any web server and thus will be accessible to any interested party. Future development should allow for the vocabulary to be directly linked to a LEXUS database and thus also allow for access right management.

Improvements on the Distributed Architecture for Assisted Annotation of Video Corpora

Rémi Dubot and Christophe Collet

Progress on automatic annotation looks attractive for the research on sign languages. Unfortunately, such tools are not easy to deploy or share. We propose a solution to uncouple the annotation software from the automatic processing module.

Such a solution requires many developments: design of a network stack supporting the architecture, production of a video server handling trust policies, standardization of annotation encoding.

In this article, we detail the choices made to implement this architecture.

Semi-Automatic Annotation of Semantic Relations in a Swiss German Sign Language Lexicon

Sarah Ebling, Katja Tissi and Martin Volk

We propose an approach to semi-automatically obtaining semantic relations in Swiss German Sign Language (Deutschschweizerische Gebärdensprache, DSGS). We use a set of keywords including the gloss to represent each sign. We apply GermaNet, a lexicographic reference database for German annotated with semantic relations. The results show that approximately 60% of the semantic relations found for the German keywords associated with 9000 entries of a DSGS lexicon also apply for DSGS. We use the semantic relations to extract sub-types of the same type within the concept of double glossing (Konrad 2011). We were able to extract 53 sub-type pairs.

Sign Language Technologies and Resources of the Dicta-Sign Project

Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos and François Lefebvre-Albaret

Here we present the outcomes of Dicta-Sign FP7-ICT project. Dicta-Sign researched ways to enable communication between Deaf individuals through the development of human-computer interfaces (HCI) for Deaf users, by means of Sign Language. It has researched and developed recognition and synthesis engines for sign languages (SLs) that have brought sign recognition and generation technologies significantly closer to authentic signing. In this context, Dicta-Sign has developed several technologies demonstrated via a sign language aware Web 2.0, combining work from the fields of sign language recognition, sign language animation via avatars and sign language resources and language models development, with the goal of allowing Deaf users to make, edit, and review avatar-based sign language contributions online, similar to the way people nowadays make text-based contributions on the Web.

Towards Tagging of Multi-Sign Lexemes and other Multi-Unit Structures

Thomas Hanke, Susanne König, Reiner Konrad and Gabriele Langer

With the building of larger sign language corpora tagging, handling and analysing large amounts of data reach a new level of complexity. Efficiency and interpersonal consistency in tagging are relevant issues as well as procedures and structures to identify and tag relevant linguistic units and structures beyond and above the manual sign level. We present and discuss problems and possible solution approaches (focussing on the working environment of iLex) of how to deal with multi-unit structures and more specifically multi-sign lexemes in annotation and lexicon building.

From Form to Function. A Database Approach to Handle Lexicon Building and Spotting Token Forms in Sign Languages

Reiner Konrad, Thomas Hanke, Susanne König, Gabriele Langer, Silke Matthes, Rie Nishio and Anja Regen

Using a database with type entries that are linked to token tags in transcripts has the advantage that consistency in lemmatising is not depending on ID-glosses. In iLex types are organised in different levels. The type hierarchy allows for analysing form, iconic value, and conventionalised meanings of a sign (sub-types). Tokens can be linked either to types or sub-types.

We expanded this structure for modelling sign inflection and modification as well as phonological variation. Differences between token and type form are grouped by features, called qualifiers, and specified by feature values (vocabularies). Built-in qualifiers allow for spotting the form difference when lemmatising. This facilitates lemma revision and helps to get a clear picture of how inflection,

modification, or phonological variation is distributed among lexical signs. This is also a strong indicator for further POS tagging. In the long term this approach will extend the lexical database from citation-form closer to full-form.
The paper will explain the type hierarchy and introduce the qualifiers used up-to-date. Further on the handling and how the data are displayed will be illustrated. As we report work in progress in the context of the DGS corpus project, the modelling is far from complete.

A Conceptual Approach in Sign Language Classification for Concepts Network

Cedric Moreau

Most websites presuppose a conceptual equivalence between a written word and a sign. In such tools, signs which do not have strict written equivalent lexicons cannot be found. The collaborative website OCELLES project LSF/French tries to give the opportunity to obtain several signs for a unique concept, with the possibility of uploading a sign without being constrained by written language. Although word checking in a written text is quite easy, this is not the case for sign checking in a video. Today studies are carried out in the field of gesture recognition, but all the sign language linguistic parameters cannot be considered as such. Indeed, they have to be used simultaneously during communication interactions. Our approach based upon the semiological Cuxac model (Cuxac, 2000) and Thom morphogenesis theory (Thom, 1973), could help to find a sign in a sign dictionary without using any written language.

A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface

Carol Neidle and Christian Vogler

A significant obstacle to broad utilization of corpora is the difficulty in gaining access to the specific subsets of data and annotations that may be relevant for particular types of research. With that in mind, we have developed a web-based Data Access Interface (DAI), to provide access to the expanding datasets of the American Sign Language Linguistic Research Project (ASLLRP). The DAI facilitates browsing the corpora, viewing videos and annotations, searching for phenomena of interest, and downloading selected materials from the website. The web interface, compared to providing videos and annotation files off-line, also greatly increases access by people that have no prior experience in working with linguistic annotation tools, and it opens the door to integrating the data with third-party applications on the desktop and in the mobile space. In this paper we give an overview of the available videos, annotations, and search functionality of the DAI, as well as plans for future enhancements. We also summarize best practices and key lessons learned that are crucial to the success of similar projects.

A Proposal for Making Corpora More Accessible for Synthesis: A Case Study Involving Pointing and Agreement Verbs

Rosalee Wolfe, John C. McDonald, Jorge Toro and Jerry Schnepp

Sign language corpora serve many purposes, including linguistic analysis, curation of endangered languages, and evaluation of linguistic theories. They also have the potential to serve as an invaluable resource for improving sign language synthesis. Making corpora more accessible for synthesis requires geometric as well as linguistic data. We explore alternate approaches and analyze the tradeoffs for the case of synthesizing indexing and agreement verbs. We conclude with a series of questions exploring the feasibility of utilizing corpora for synthesis.

Session C: Resources

Sunday 27 May, 14:00 – 16:00
Chairperson: Jette Kristoffersen Poster Session

SIGNSPEAK Project Tools: A Way to Improve the Communication Bridge between Signer and Hearing Communities

Javier Caminero, Mari Carmen Rodriguez-Gancedo, Alvaro Hernandez-Trapote and Beatriz Lopez-Mencia

The SIGNSPEAK project is aimed at developing a novel scientific approach for improving the communication between signer and hearing communities. In this way, SIGNSPEAK technology captures the video information from the signer and converts it into text. To do that, SIGNSPEAK consortium has devoted great efforts to the creation and annotation of the RWTH-Phoenix corpus. Based on it, a multimodal processing of the captured video is carried out and the resultant sign sequence is translated into natural language. Afterwards, the intended message could be communicated to hearing-able people using a text-to-speech (TTS) engine. In the reverse way, speech from hearing-able people would be transformed into text using Automatic Speech Recognition (ASR) and then the text would be processed by virtual avatars able to compose the suitable sign sequence. In SIGNSPEAK project, scientific and usability approaches have been combined to go beyond the state-of-the-art and contributing to suppress barriers between signer and hearing communities. In this work, a special stress was put in the development of a prototype and also, in setting of the grounds for future real industrial applications.

From Corpus to Lexicon: The Creation of ID-Glosses for the Corpus NGT

Onno Crasborn and Anne de Meijer

When glossing of the Corpus NGT started in 2007, there was no lexicon at our disposal to base ID-glosses on. Semantic labels were used without ensuring a constant relationship between sign form and gloss. This is currently being repaired by creating a lexicon from scratch alongside with the creation of new annotations. This substantial task is still in progress, but promises to lead to several new research avenues for the future. The current paper describes some of the choices that were made in the process, and specifies some of the glossing conventions that were used.

A GSL Continuous Phrase Corpus: Design and Acquisition

Athanasia-Lida Dimou, Vassilis Pitsikalis, Theodoros Goulas, Stavros Theodorakis, Panagiotis Karioris, Michalis Pissaris, Stavroula-Evita Fotinea, Eleni Efthimiou and Petros Maragos

The corpus presented in this article is composed of a limited number of Greek Sign Language (GSL) sentences and was created in order to provide additional data to the already obtained corpus during the first year of the Dicta-Sign project (Matthes et al., 2010). More specifically this corpus intended to serve as the ground upon which a significant part of the recognition process would be tested and evaluated, more precisely, the continuous sign language recognition algorithms developed in the project.

Given the targeted nature of this corpus we present here the constraints as well as the procedure followed in order to obtain it.

The procedure followed for the creation of this corpus, consists of its linguistic design and validation, the studio and hardware acquisition configuration, the implementation and supervision of the acquisition itself and the post-processing and annotation of the obtained data in order to release the set of usable annotated resources. The specific GSL phrase corpus forms the basis for machine learning and training to serve experimentation in the domain of continuous sign language processing and recognition.

A Study on Qualification/Naming Structures in Sign Languages

Michael Filhol and Annelies Braffort

In the prospect of animating virtual signers, this article addresses the issue of representing Sign, in particular on levels not restricted to the language lexicon. In order to choose and design a suitable model, we illustrate the main steps of our corpus-based methodology for linguistic structure identification and formal description with the example of a specific structure we have named "qualification/naming". We also discuss its similarity and difference with other Sign properties described in the literature such as compound signs. Consequently we explain our choice for a description model that does not separate lexicon and grammar in two disjoint levels for virtual signer input.

Experiences Collecting Motion Capture Data on Continuous Signing

Tommi Jantunen, Birgitta Burger, Danny De Weerdt, Irja Seilola and Tuija Wainio

This paper describes some of the experiences the authors have had collecting continuous motion capture data on Finnish Sign Language in the motion capture laboratory of the Department of Music at the University of Jyväskylä, Finland. Monologue and dialogue data have been recorded with an eight-camera optical motion capture system by tracking, at a frame rate of 120 Hz, the three-dimensional locations of small ball-shaped reflective markers attached to the signer's hands, arms, head, and torso. The main question from the point of view of data recording concerns marker placement, while the main themes discussed concerning data processing include gap-filling (i.e. the process of interpolating the information of missing frames on the basis of surrounding frames) and the importing of data into ELAN for subsequent segmentation (e.g. into signs and sentences). The paper will also demonstrate how the authors have analyzed the continuous motion capture data from the kinematic perspective.

Towards Russian Sign Language Synthesizer: Lexical Level

Alexey Karpov and Miloš Železný

In this paper, we present a survey of existing Russian sign language electronic and printed resources and dictionaries. The problem of differences in dialects of Russian sign language used in various local communities of Russia and some other CIS countries is discussed in the paper. Also the first version of a computer system for synthesis of elements of Russian sign language (signed Russian and fingerspelling) is presented in the given paper. It is a universal multi-modal synthesizer both for Russian spoken language and signed Russian that is based on a model of animated 3D signing avatar. The proposed system inputs data in the text form and converts them into the audio-visual modality, synchronizing visual manual gestures and articulation with audio speech signal. Generated audio-visual signed Russian speech and spoken language is a fusion of dynamic gestures shown by the avatar's both hands, lip movements articulating words and auditory speech, so the multimodal output is available both for the deaf and hearing-able people.

A Colorful First Glance at Data on Regional Variation Extracted from the DGS-Corpus: With a Focus on Procedures

Gabriele Langer

In this work in progress procedures for analyzing and displaying distributional patterns of sign variants have been developed and tested on data for color signs elicited by the DGS Corpus Project. The data for this preliminary study were elicited as isolated signs and have been made accessible through spot annotations in iLex. The annotations had not been lemma revised but nevertheless revealed some interesting insights. Several color signs exhibited a high degree of variation. The distributional maps showed that a number of signs were mainly used in certain regions and thus provided evidence on dialectal differences within DGS. The relevant information necessary to generate distributional maps have been directly extracted via SQL-statements from the corpus and fed into R. The approach is data driven. The distributional maps show either the distribution of one sign form (variant) or of several different variants in relation to each other. Analyses of regional distribution as displayed by the distributional maps may support the annotation and lemma revision process and are a valuable basis for a lexicographical description of signs and their use as needed for compiling dictionary entries. A refined procedure to take multiple regional influences on informants into account for analysis is proposed.

CUNY American Sign Language Motion-Capture Corpus: First Release

Pengfei Lu and Matt Huenerfauth

We are in the middle of a 5-year study to collect, annotate, and analyze an ASL motion-capture corpus of multi-sentential discourse. Now we are ready to release to the research community the first sub-portion of our corpus that has been checked for quality. This paper describes the recording and annotation procedure of our released corpus to enable researchers to determine if it would benefit their work. A focus of the collection process was the identification and use of prompting strategies for eliciting single-signer multi-sentential ASL discourse that maximizes the use of pronominal spatial reference yet minimizes the use of classifier predicates. The annotation of the corpus includes details about the establishment and use of pronominal spatial reference points in space. Using this data, we are seeking computational models of the referential use of signing space and of spatially inflected verb forms for use in American Sign Language (ASL) animations, which have accessibility applications for deaf users.

Dicta-Sign – Building a Multilingual Sign Language Corpus

Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert and Eva Safar

This paper presents the multilingual corpus of four European sign languages compiled in the framework of the Dicta-Sign project. Dicta-Sign researched ways to enable communication between Deaf individuals through the development of human-computer interfaces (HCI) for Deaf users, by means of sign language. Sign language resources were compiled to inform progress in the other research areas within the project, especially video recognition of signs, sign-to-sign translation, linguistic modelling, and sign generation. The aim for the corpus data collection was to achieve as high a level of naturalness as possible with semi-spontaneous utterances under lab conditions. At the same time the elicited data were supposed to be semantically close enough to be comparable both across individual informants and for all four sign languages. The sign language data were annotated using iLex and are now made available via a web portal that allows for different access options to the data.

Sign Language Resources in Sweden: Dictionary and Corpus

Johanna Mesch, Lars Wallin and Thomas Björkstrand

Sign language resources are necessary tools for adequately serving the needs of learners, teachers and researchers of signed languages. Among these resources, the Swedish Sign Language Dictionary was begun in 2008 and has been in development ever since. Today, it has approximately 8,000 sign entries. The Swedish Sign Language Corpus is also an important resource, but it is of a very different kind than the dictionary. Compiled during the years 2009–2011, the corpus consists of video recorded conversations among 42 informants aged between 20 and 82, from three separate regions in Sweden. With 14 % of the corpus having been annotated with glosses for signs, it comprises total of approximately 3,600 different signs occurring about 25,500 times (tokens) in the 42 annotated sign language discourses/video files. As these two resources sprang from different starting points, they are independent from each other; however, in the late phases of building the corpus the importance of combining work from the two became evident. This presentation will show the development of these two resources and the advantages of combining them.

English-ASL Gloss Parallel Corpus 2012: ASLG-PC12

Achraf Othman and Mohamed Jemni

A serious problem facing the Community for researchers in the field of sign language is the absence of a large parallel corpus for signs language. The ASLG-PC12 project proposes a rule-based approach for building big parallel corpus between English written texts and American Sign Language Gloss. We present a novel algorithm which transforms an English part-of-speech sentence to ASL gloss. This project was started in the beginning of 2011, a part of the project WebSign, and it offers today a corpus containing more than one hundred million pairs of sentences between English and ASL gloss. It is available online for free in order to develop and design new algorithms and theories for American Sign Language processing, for example statistical machine translation and any related fields. In this paper, we present tasks for generating ASL sentences from the corpus Gutenberg Project that contains only English written texts.

Compiling the Slovene Sign Language Corpus

Špela Vintar, Boštjan Jerko and Marjetka Kulovec

We report on the project of compiling the first corpus of the Slovene Sign Language. The paper describes the procedures of data collection, the decisions regarding informant selection and plans for transcription and annotation. We outline the particularities of the Slovene situation, especially the high variability of the language, issues concerning language competence and the attitutes of the deaf community towards such data collection. At the time of writing, the data collection stage is nearly finished with over 70 persons recorded, and trancriptions with iLex are underway. The aim of the project is to use the corpus for explorations into the grammatical properties of SSL.

Session D: Issues in the Construction of Sign Corpora

Sunday 27 May, 16:30 – 19:00
Chairperson: Stavroula-Evita Fotinea

Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus

Carol Neidle, Ashwin Thangali and Stan Sclaroff

The American Sign Language Lexicon Video Dataset (ASLLVD) consists of videos of >3,300 ASL signs in citation form, each produced by 1-6 native ASL signers, for a total of almost 9,800 tokens. This dataset, including multiple synchronized videos showing the signing from different angles, will be shared publicly once the linguistic annotations and verifications are complete. Linguistic annotations include gloss labels, sign start and end time codes, start and end handshape labels for both hands, morphological and articulatory classifications of sign type. For compound signs, the dataset includes annotations for each morpheme. To facilitate computer vision-based sign language recognition, the dataset also includes numeric ID labels for sign variants, video sequences in uncompressed-raw format, camera calibration sequences, and software for skin region extraction. We discuss here some of the challenges involved in the linguistic annotations and categorizations. We also report an example computer vision application that leverages the ASLLVD: the formulation employs a HandShapes Bayesian Network (HSBN), which models the transition probabilities between start and end handshapes in monomorphemic lexical signs. Further details and statistics for the ASLLVD dataset, as well as information about annotation conventions, are available from http://www.bu.edu/asllrp/lexicon.

SignWiki – An Experiment in Creating a User-based Corpus

Sonja Erlenkamp and Olle Eriksen

In comparison to other signed languages, Norwegian Sign Language (NTS) is not well researched and documented while at the same time the need for documentation of NTS in a corpus based dictionary has been apparent to the field for quite a while. Despite some high quality applications to raise funding for corpus work, the field in Norway has not succeeded to gain enough understanding in governmental research funding institutions for the need of a corpus based dictionary, mainly because of the rather small population of NTS users. As a result, a new approach is used by involving the NTS community to create a database of signs, including their use, distribution and as far as possible other metadata. Tegnwiki (=Signwiki) is a first attempt at creating a user-based database of NTS by allowing users to contribute videos and information on isolated signs on a Wikiplatform. Like Wikipedia, the SignWiki will be open accessible, but administered by a group of experts. Obviously a SignWiki cannot replace a scientific corpus. But if this experiment is successful it might be a good starting point for countries with no or little funding for corpus projects where involvement of users is the key factor.

Where Does a Sign Start and End? Segmentation of Continuous Signing

Thomas Hanke, Silke Matthes, Anja Regen and Satu Worseck

There are two basic approaches how to segment continuous signing into individual signs:

- A sign starts where the preceding one ends (i.e. fluent signing means there are no gaps between signs)
- Transitional movements between signs do not count as part of either sign. Therefore, usually there are gaps between two signs during which the articulators move from the end of one sign to the beginning of the next.

Both approaches have their pros and cons. However, in the context of the DGS Corpus and the Dicta-Sign project the second approach offers advantages for the subsequent processing. Here we investigate how sensitive this approach is with respect to higher video frame rates.

Transcribing and Evaluating Language Skills of Deaf Children in a Multimodal and Bilingual Way: the Sensitive Issue of the Gesture/Signs Dynamics

Isabelle Estève

Transcribing and evaluating the narrative productions of 6 to 12 year-olds deaf children in their multimodal and bilingual dimensions confront us to the central question of gestures/signs distinction. This paper aims to discuss how the narrative skills of 30 deaf children schooled in different education settings – oralist, bilingual and “mixed” – led us to create transcription/annotation tools in ELAN allowing to take into account the dynamics between verbal and non-verbal material involving especially within the gestural modality. We will focus on two central points of our reflections. How to delimit productions in units into taking into account the semiotic and the structural dynamics aspects of production? How to describe and categorize the gestural processes non systematized in a linguistic form to report the developmental dynamics?

Sign Language Documentation in the Asia-Pacific Region: A Deaf-centred approach

Felix Sze, James Woodward, Gladys Tang, Jafi Lee, Ka-Yiu Cheng and Joe Ma

In this paper, we would like to share our experience in training up Deaf individuals from the Asian-Pacific countries to compile sign language dictionaries and conduct sign language research through the ‘Asia-Pacific Sign Linguistics Research and Training Program’ at the Chinese University of Hong Kong. The program, fully funded by the Nippon Foundation, is a multi-country, multi-phase project which aims at nurturing Deaf people to become sign language researchers through a series of credit-bearing training programs at the diploma and higher diploma levels. The training covers three major areas: Sign Linguistics, Sign Language Teaching and English Literacy. One important part of the training involves the production of sample dictionaries of the Deaf trainees’ own sign languages. To confirm the dictionary entries, the Deaf trainees conduct surveys in the Deaf communities in their home countries from time to time and as a result a substantial amount of lexical variants have been collected. An online database, called the Asian SignBank, is now being developed to house these lexical data and facilitate further research. Apart from basic search functions, the SignBank also incorporates detailed phonetic features of individual signs and a materials-generating function which allows quicker production of dictionaries in the future.