8th Workshop on the Representation and Processing of Sign Languages:

Involving the Language Community

12 May 2018

ABSTRACTS

Editors:

Mayumi Bono, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, Johanna Mesch, Yutaka Osugi

Workshop Programme

09:00 – 10:30	On-stage Session A: *Corpora*
10:30 – 11:00	Coffee break
11:00 – 13:00	Poster Session B: *Corpus-based Research and Development*
13:00 – 14:00	Lunch break
14:00 – 16:00	Poster Session C: *Corpora and the Language Community*
16:00 – 16:30	Coffee break
16:30 – 18:00	On-stage Session D: *Involving the Language Community*

Workshop Organizers

Mayumi Bono	National Institute of Informatics, Tokyo JP
Eleni Efthimiou	Institute for Language and Speech Processing, Athens GR
Stavroula-Evita Fotinea	Institute for Language and Speech Processing, Athens GR
Thomas Hanke	Institute of German Sign Language, University of Hamburg, Hamburg DE
Julie Hochgesang	Gallaudet University, Washington US
Jette Kristoffersen	Centre for Sign Language, University College Copenhagen, Copenhagen DK
Johanna Mesch	Stockholm University, Stockholm SE
Yutaka Osugi	Tsukuba University of Technology, Tsukuba JP

Workshop Programme Committee

Mayumi Bono	National Institute of Informatics, Tokyo JP
Annelies Braffort	LIMSI/CNRS, Orsay FR
Onno Crasborn	Radboud University, Nijmegen NL
Sarah Ebling	University of Applied Sciences of Special Needs (HfH), Zurich CH
Eleni Efthimiou	Institute for Language and Speech Processing, Athens GR
Michael Filhol	CNRS–LIMSI, Université Paris-Saclay, Orsay FR
Stavroula-Evita Fotinea	Institute for Language and Speech Processing, Athens GR
Sylvie Gibet	IRISA, University of Bretagne Sud, Vannes FR
Thomas Hanke	Institute of German Sign Language, University of Hamburg, Hamburg DE
Alexis Héloir	DFKI, Saarbrücken DE
Julie Hochgesang	Gallaudet University, Washington US
Matt Huenerfauth	Rochester Institute of Technology, Rochester, NY, USA
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 Copenhagen, Copenhagen DK
John McDonald	DePaul University, Chicago US
Johanna Mesch	Stockholm University, Stockholm SE
Carol Neidle	Boston University, Boston US
Yutaka Osugi	Tsukuba University of Technology, Tsukuba JP
Christian Rathmann	Humboldt-Universität zu Berlin, Berlin DE
Rosalee Wolfe	DePaul University, Chicago US

Session A: Corpora

Saturday 12 May, 09:00 – 10:30
On-stage Session

Scalable ASL Sign Recognition using Model-based Machine Learning and Linguistically Annotated Corpora

Dimitri Metaxas, Mark Dilsizian and Carol Neidle

We report on the high success rates of our new, scalable, signer-independent, computational approach for sign recognition from monocular video, exploiting linguistically annotated ASL data sets. We recognize signs using a hybrid framework that combines state-of-the-art learning methods with features based on what is known about the linguistic composition of lexical signs. We model and recognize the sub-components of sign production, with attention to hand shape, orientation, location, motion trajectories, as well as facial features, and we combine these within a CRF framework. The effect is to make the sign recognition problem robust, scalable, and feasible with relatively smaller datasets than are required for purely data-driven methods. From a 350-sign vocabulary of isolated, citation-form lexical signs from the American Sign Language Lexicon Video Dataset (ASLLVD), including both 1- and 2-handed signs, we achieve a top-1 accuracy of 93.6% and a top-5 accuracy of 97.9%. The high probability with which we can produce 5 sign candidates that contain the correct result opens the door to potential applications, as it is reasonable to provide a sign lookup functionality that offers the user 5 possible signs, in decreasing order of likelihood, with the user then asked to select the desired sign.

Building French Sign Language Motion Capture Corpora for Signing Avatars

Sylvie Gibet

The design of traditional corpora for linguistic analysis aims to provide living representations of sign languages across deaf communities and linguistic researchers. Most of the time, the sign language data is video-recorded and then encoded in a standardized and homogenous structure for open-ended analysis (statistical or phonological studies). With such structures, sign language corpora are described and annotated into linguistic components, including phonology, morphology, and syntactic components. Conversely, motion capture (MoCap) corpora provide researchers the data necessary to carry on finer-grained studies on movement, thus allowing precise, and quantitative analysis of sign language gestures as well as sign language (SL) generation. One the one hand, motion data serves to validate and enforce existing theories on the phonologies of sign languages. By aligning temporally motion trajectories and labelled linguistic information, it thus becomes possible to study the influence of the movement articulation on the linguistic aspects of the SL, including hand configuration, hand movement, co-articulation or synchronization within intra and inter phonological channels. On the other hand, generation pertains to sign production using animated virtual characters, usually called signing avatars. Although MoCap technology presents exciting future directions for sign language studies, tightly interlinking language components and signals, it still requires high technical skills for recording, post-processing data, and there are many unresolved challenges, with the need to simultaneously record body, hand motion, facial expressions, and gaze direction. Therefore, there are still few MoCap corpora that have been developed in the field of sign language studies. Some of them are dedicated to the analysis of articulation and prosody aspects of sign languages, whereas recent interest in avatar technology has led to develop corpora associated to data-driven synthesis. This paper describes four corpora that have been designed and built in our research team. These corpora have been recorded using MoCap and video equipment, and annotated according to multi-tiers linguistic templates. Each corpus has been designed for a specific linguistic purpose and is dedicated to data-driven synthesis, by

replacing signs or groups of signs, by composing phonetic or phonological components, or by altering prosody in the produced sign language utterances.

From Design and Collection to Annotation of a Learner Corpus of Sign Language

Johanna Mesch and Krister Schönström

This paper aims to present part of the project "From Speech to Sign – learning Swedish Sign Language as a second language" which include a learner corpus that is based on data produced by hearing adult L2 signers. The paper describes the design of corpus building and the collection of data for the Corpus in Swedish Sign Language as a Second Language (SSLC-L2). Another component of ongoing work is the creation of a specialized annotation scheme for SSLC-L2, one that differs somewhat from the annotation work in Swedish Sign Language Corpus (SSLC), where the data is based on performance by L1 signers. Also, we will account for and discuss the methodology used to annotate L2 structures.

Session B: Corpus-based Research and Development

Saturday 12 May, 11:00 – 13:00
Poster Session

Modeling and Predicting the Location of Pauses for the Generation of Animations of American Sign Language

Sedeeq Al-khazraji, Sushant Kafle and Matt Huenerfauth

Adding American Sign Language (ASL) animation to websites can improve information access for people who are deaf with low levels of English literacy. Given a script representing the sequence of ASL signs, we must generate an animation, but a challenge is selecting accurate speed and timing for the resulting animation. In this work, we analyzed motion-capture data recorded from human ASL signers to model the realistic timing of ASL movements, with a focus on where to insert prosodic breaks (pauses), based on the sentence syntax and other features. Our methodology includes extracting data from a pre-existing ASL corpus at our lab, selecting suitable features, and building machine learning models to predict where to insert pauses. We evaluated our model using cross-validation and compared various subsets of features. Our model had 80% accuracy at predicting pause locations, out-performing a baseline model on this task.

Augmenting Sparse Corpora for Enhanced Sign Language Recognition and Generation

Heike Brock, Juliette Rengot and Kazuhiro Nakadai

The collection of signed utterances for recognition and generation of Sign Language (SL) is a costly and labor-intensive task. As a result, SL corpora are usually considerably smaller than their spoken language or image data counterparts. This is problematic, since the accuracy and applicability of a neural network depends largely on the quality and amount of its underlying training data. Common data augmentation strategies to increase the number of available training data are usually not applicable to the spatially and temporally constrained motion sequences of a SL corpus. In this paper, we therefore discuss possible data manipulation methods on the base of a collection of motion-captured SL sentence expressions. Evaluation of differently trained network architectures shows a significant reduction of overfitting by inclusion of the augmented data. Simultaneously, the accuracy of both sign recognition and generation was improved, indicating that the proposed data augmentation methods are beneficial for constrained and sparse data sets.

Communication Across Sensorial Divides – A Proposed Community Sourced Corpus of Everyday Interaction between Deaf Signers and Hearing Nonsigners

Mio Cibulka

While research on conversation in signed and spoken languages has been flourishing, research on their intersection is scarce. This paper presents an ongoing project that gathers and analyses video data from deaf people's everyday interaction with hearing nonsigners and considers possibilities of involving the communication community that is at its centre and participant empowerment. The scope is to investigate the organisation and structure of communication in which linguistic resources are less accessible and in which social meaning tends to emerge from the interactants' online analysis of the local context (e.g., spatial environment, bodily configurations and movement of the interactants).

SiLOrB and Signotate: A Proposal for Lexicography and Corpus-building via the Transcription, Annotation, and Writing of Signs

Brenda Clark and Greg Clark

This paper proposes a system of standardized transcription and orthographic representation for sign languages (Sign Language Orthography Builder) with a corresponding text-based corpus-building and annotation tool (Signotate). The transcription system aims to be analogous to IPA in using ASCII characters as a standardized way to represent the phonetic aspects of any sign, and the writing system aims to be transparent and easily readable, using pictographic symbols which combine to create a 'signer' in front of the reader. The proposed software can be used to convert transcriptions to written signs, and to create annotated corpora or lexicons. Its text-based human- and machine-readable format gives a user the ability to search large quantities of data for a variety of features and contributes to sources, such as dictionaries and transcription corpora.

The POLYTROPON Parallel Corpus

Eleni Efthimiou, Kiki Vasilaki, Stavroula-Evita Fotinea, Anna Vacalopoulou, Theodore Goulas and Athanasia-Lida Dimou

Here we present the POLYTROPON parallel corpus for the language pair Greek Sign Language (GSL) – Greek, which is created and annotated aiming to serve as a golden corpus available to the community of SL technologies for experimentation with various approaches to SL processing, focusing on machine learning for SL recognition, machine translation (MT) and information retrieval. The corpus volume incorporates 3653 clauses in three repetitions each, captured in front view by means of one HD and one Kinect camera. Corpus annotation has allowed to extract initial features sets with the aim to reach a GSL level of abstraction close to the one currently available for Greek language representations, exploiting the inherent characteristics of the language in view of applying initial deep learning experiments on GSL data, where both words and signs may be represented as vectors of characteristics which allow dependency tree structure representations of input text and signed clauses as those created by the use of Tree Editor TrEd 2.0.

Extending the AZee-Paula Shortcuts to Enable Natural Proform Synthesis

Michael Filhol and John McDonald

Proform structures such as classifier predicates have traditionally challenged Sign Language (SL) synthesis systems, particularly in respect to the production of smooth natural motion. To address this issue a synthesizer must necessarily leverage a structured linguistic model for such constructs to specify the linguistic constraints, and also an animation system that is able to provide natural avatar motion within the confines of those constraints. The proposed system bridges two existing technologies, taking advantage of the ability of AZee to encode both the form and functional linguistic aspects of the proform movements and on the Paula avatar system to provide convincing human motion. The system extends a previous principle that more natural motion arises from leveraging knowledge of larger structures in the linguistic description.

Building the ASL Signbank: Lemmatization Principles for ASL

Julie Hochgesang, Onno Crasborn and Diane Lillo-Martin

Following the example of other sign language researchers, we are creating a Signbank, a usage-based lexical database, to maintain consistent and systematic annotation information for American Sign Language (ASL). This tool, which will be available to the public, is currently being used in conjunction with an on-going effort to prepare corpora of sign language acquisition to share with the research community. This paper will briefly report on the development of the ASL Signbank, focusing on the adopted lemmatization principles. Lemmatization of ASL signs has never been done on a scale like this before - one that has been continually refreshed by actual usage data.

Development of an "Integrative System for Korean Sign Language Resources"

Sung-Eun Hong, Seongok Won, Il Heo and Hyunhwa Lee

In 2015, the KSL Corpus Project started to create a linguistic corpus of the Korean Sign Language (KSL). The collected data contains about 90 hours of sign language videos. Almost 17 hours of this sign language data has been annotated in ELAN, a professional annotation tool developed by the Max-Planck-Institute of Psycholinguistics in the Netherlands. In the first phase of annotation the research project faced three major difficulties. First there was no lexicon or lexical database available that means the annotators had to list the used sign types and link them with video clips showing the sign type. Second, having numerous annotators it was a challenge to manage and distribute the hundreds of movies and ELAN files. Third it was very difficult to control the quality of the annotation. In order to solve these problems the "Integrative System for Korean Sign Language Resources" was developed. This system administrates the signed movies and annotations files and also keeps track of the lexical database. Since all annotation files are uploaded into the system, the system is also able to manipulate the ELAN files. For example, tags are overwritten in the annotation when the name of the type has changed.

Quotation in Russian Sign Language: A Corpus Study

Vadim Kimmelman and Evegniia Khristoforova

We studied how quotation is expressed in naturalistic discourse in Russian Sign Language (RSL). We studied a sub-corpus of the online corpus of RSL containing narratives by eleven signers from Moscow. We identified 341 instances of quotation, including reported speech and reported thoughts. We annotated syntactic, semantic, and prosodic properties of the found instances of quotation. We found out that quotative constructions in RSL have the same basic structure as

similar constructions in other spoken and signed languages. Furthermore, similarly to quotation in other sign languages, quotation in RSL can be marked by head and/or body movement and change in eye gaze direction. However, all of these markers are clearly optional, and a considerable number of examples do not include any of these markers. Furthermore, we found that, judging by the behavior of indexicals, RSL narratives in our dataset have a very strong preference for using direct speech. We discuss theoretical implications of the RSL data to the theory of quotation in sign languages.

What Corpus-based Research on Negation in Auslan and PJM Tells Us about Building and Using Sign Language Corpora

Anna Kuder, Joanna Filipczak, Piotr Mostowski, Paweł Rutkowski and Trevor Johnston

In this paper, we would like to discuss our current work on negation in Auslan (Australian Sign Language) and PJM (Polish Sign Language, polski język migowy) as an example of experience in using sign language corpus data for research purposes. We describe how we prepared the data for two detailed empirical studies, given similarities and differences between the Australian and Polish corpus projects. We present our findings on negation in both languages, which turn out to be surprisingly similar. At the same time, what the two corpus studies show seems to be quite different from many previous descriptions of sign language negation found in the literature. Some remarks on how to effectively plan and carry out the annotation process of sign language texts are outlined at the end of the present paper, as they might be helpful to other researchers working on designing a corpus. Our work leads to two main conclusions: (1) in many cases, usage data may not be easily reconciled with intuitions and assumptions about how sign languages function and what their grammatical characteristics are like, (2) in order to obtain representative and reliable data from large-scale corpora one needs to plan and carry out the annotation process very thoroughly.

Queries and Views in iLex to Support Corpus-based Lexicographic Work on German Sign Language (DGS)

Gabriele Langer, Anke Müller and Sabrina Wähl

In the DGS-Korpus project the corpus is being used as the basis for lexicographic descriptions of signs in dictionary entries. In this process the lexicographers start from the data and type entry structures as found in the annotation database. While preparing a dictionary entry much of the work consists of manually going through a number of single tokens viewing the original data and available annotations. Findings are then categorised and summarised. However, a number of decisions and descriptions are also supported by pre-defined searches and views on the data. Supported areas include lexicographic lemmatisation (lemma sign establishment), selection of citation forms and variants, grammatical behaviour of signs, collocational patterns of use, regional distribution patterns and distribution of lexical or formational variants over different age groups. While we are still in the process of exploring the possibilities of a sign language corpus for lexicography, searches and views that have proven useful for our work are exemplified in this paper with regard to dictionary entries.

Per Channel Automatic Annotation of Sign Language Motion Capture Data

Lucie Naert, Clément Reverdy, Caroline Larboulette and Sylvie Gibet

Manual annotation is an expensive and time consuming task partly due to the high number of linguistic channels that usually compose sign language data. In this paper, we propose to automatize

the annotation of sign language motion capture data by processing each channel separately. Motion features (such as distances between joints or facial descriptors) that take advantage of the 3D nature of motion capture data and the specificity of the channel are computed in order to (i) segment and (ii) label the sign language data. Two methods of automatic annotation of French Sign Language utterances using similar processes are developed. The first one describes the automatic annotation of thirty-two hand configurations while the second method describes the annotation of facial expressions using a closed vocabulary of seven expressions. Results for the two methods are then presented and discussed.

NEW Shared & Interconnected ASL Resources: SignStream® 3 Software; DAI 2 for Web Access to Linguistically Annotated Video Corpora; and a Sign Bank

Carol Neidle, Augustine Opoku, Gregory Dimitriadis and Dimitri Metaxas

2017 marked the release of a new version of SignStream® software, designed to facilitate linguistic analysis of ASL video. SignStream® provides an intuitive interface for labeling and time-aligning manual and non-manual components of the signing. Version 3 has many new features. For example, it enables representation of morpho-phonological information, including display of handshapes. An expanding ASL video corpus, annotated through use of SignStream®, is shared publicly on the Web. This corpus (video plus annotations) is Web-accessible—browsable, searchable, and downloadable—thanks to a new, improved version of our Data Access Interface: DAI 2. DAI 2 also offers Web access to a brand new Sign Bank, containing about 10,000 examples of about 3,000 distinct signs, as produced by up to 9 different ASL signers. This Sign Bank is also directly accessible from within SignStream®, thereby boosting the efficiency and consistency of annotation; new items can also be added to the Sign Bank. Soon to be integrated into SignStream® 3 and DAI 2 are visualizations of computer-generated analyses of the video: graphical display of eyebrow height, eye aperture, and head position. These resources are publicly available, for linguistic and computational research and for those who use or study ASL.

The LESCO Corpus. Data for the Description of Costa Rican Sign Language

Alejandro Oviedo and Christian Ramírez Valerio

LESCO is the most widely used sign language among Deaf people in Costa Rica (Woodward 1992). There are no precise figures available on the number of LESCO users, who live mostly in urban areas of the Central Valley of the country. Between 2010 and 2013 the Costa Rican government funded a project for a first linguistic description of LESCO, as a step towards recognition of the rights of Deaf people. The study of LESCO was based on the Corpus LESCO, a group of transcribed videos collected from Deaf signers from the main cities of the country along 2011. The project was carried on by a group consisted of five Deaf native LESCO-users and a hearing person with a good command of this sign language. A series of interviews were done over several months throughout the country allowing a pre-selection of 102 potential informants (all of them attesting a relative early LESCO acquisition, frequent use of LESCO, high degree of hearing-impairment, etc.). These people were video-recorded and so nearly 200 video-files (over 2000 minutes of footage) were obtained. Films included induced stories, life stories, free dialogues and interviews. For an initial description of the language, a selection of 44 files was transcribed on the basis of ELAN. Variants of each sign were identified and assigned to a particular lexeme. This process allowed the definition of more than 1,500 lemmas (Johnston 2010) from a total of around 14,000 lexical occurrences. The Corpus LESCO underpinned the construction of a basic dictionary (1,100 entries) and the drafting of a basic descriptive grammar of this sign language. Both dictionary and grammar are available online since the beginning of 2014 (www.cenarec-lesco.org).

These works are the second corpus-based descriptions of a signed language in Spanish speaking Latin America. A previous experienced was carried of in Colombia between 2000 and 2005 (Oviedo 2001, CyC 2005). The initial project did not include the extension of the corpus. Both the Corpus LESCO and the rest of videos collected during the project are archived by the institution that administered the project in Costa Rica. The poster offers details about the process of building up the corpus and about its main characteristics.

Recognizing Non-manual Signals in Filipino Sign Language

Joanna Pauline Rivera and Clement Ong

Filipino Sign Language (FSL) is a multi-modal language that is composed of manual signlas and non-manual signals. Very minimal research is done regarding non-manual signals (Martinez and Cabalfin, 2008) despite the fact that non-manual signals play a significant role in conversations as it can be mixed freely with manual signals (Cabalfin et al., 2012). For other Sign Languages, there have been numerous researches regarding non-manual; however, most of these focused on the semantic and/or lexical functions only. Research on facial expressions in sign language that convey emotions or feelings and degrees of adjectives is very minimal. In this research, an analysis and recognition of non-manual signals in Filipino Sign Language are performed. The non-manual signals included are Types of Sentences (i.e. Statement, Question, Exclamation), Degrees of Adjectives (i.e. Absence, Presence, High Presence), and Emotions (i.e. Happy, Sad, Fast-approaching danger, stationary danger). The corpus was built with the help of the FSL Deaf Professors, and the 5 Deaf participants who signed 5 sentences for each of the types in front of Microsoft Kinect sensor. Genetic Algorithm is applied for the feature selection, while Artificial Neural Network and Support Vector Machine is applied for classification.

Depicting Signs and Different Text Genres: Preliminary Observations in the Corpus of Finnish Sign Language

Ritva Takkinen, Jarkko Keränen and Juhana Salonen

In this article we first discuss the different kinds of signs occurring in sign languages and then concentrate on depicting signs, especially on their classification in Finnish Sign Language. Then we briefly describe the corpora of Finland's sign languages (CFINSL). The actual study concerns the occurrences of depicting signs in CFINSL in different text genres, introductions, narratives and free discussions. Depicting signs occurred most frequently in narratives, second most frequently in discussions and least frequently in introductions. The most frequent depicting signs in all genres were those that depicted the whole entity moving or being located. The second most frequent were those signs that expressed the handling of entities. The least frequent depicting signs were those with size- and shape-tracing handshapes. The proportion of depicting signs of all the signs in each genre was 17.9% in the narratives, 2.9% in the discussions and 2.2% in the introductions. In order to deepen the analysis, depicting signs will have to be investigated from the perspective of movement types and the use of one or two hands.

Session C: Corpora and the Language Community

Saturday 12 May, 14:00 – 16:00
Poster Session

Tactile Japanese Sign Language and Finger Braille: An Example of Data Collection for Minority Languages in Japan

Mayumi Bono, Rui Sakaida, Ryosaku Makino, Tomohiro Okada, Kouhei Kikuchi, Mio Cibulka, Louisa Willoughby, Shimako Iwasaki and Satoshi Fukushima

We recorded data on deafblind people in Japan. In this filming project, we found that Japanese deafblind people use different communication methods, tactile Japanese sign language and finger braille, depending on their hearing ability and eyesight. Tactile sign language is normally used by those who were born deaf or lost their hearing at an early age and then lost their sight after acquiring a sign language. These people are known as deaf-based deafblind (D-deafblind). Finger braille is popular in Japan, but largely unknown elsewhere. It is normally used by those who were born blind or lost their sight at an early age and subsequently lost their hearing after learning how to produce speech using their throat and mouth. These people are known as blind-based deafblind (B-deafblind hereafter). This paper introduces our filming project; the ways of data collection, translation and annotation. In addition, we show our preliminary observations using our data sets to clarify the important fact that we should collect their interactions at this moment. The data show how their interactions have already become established and sophisticated in their communities. We discuss how our filming project will contribute to the deafblind community in Japan.

Terminology Enrichment through Crowd Sourcing at PYLES Platform

Eleni Efthimiou, Stavroula-Evita Fotinea, Panos Kakoulidis and Theodore Goulas

The Information System PYLES is a management system for on-line lessons, designed to support accesible asynchronous e-learning, addressing learning needs of students with various communication capabilities and needs at the Technological Educational Institute of Athens (TEI-A). It, thus, exploits both uptodate assistive technology software and content in various forms. This platform has been used as the basis for the development of an active repository of multimodal educational resources, also incorporating a terminology lexicon for the Greek Sign Language (GSL) and a general purpose dictionary of GSL. The platform provides advanced customization options according to user needs but also a collaborative environment for the support of teaching and learning processes. The information system (http://eclassamea.teiath.gr/) is built on the open code platform 'Open eClass' (http://www.openeclass.org/), a free e-learning platform that it actually enriches with tools and functionalities which allow extended accessibility regarding both the environment and the educational content. Regarding customization to serve GSL signers' needs, the platform incorporates: Selected lesson presentations in GSL on the basis of deaf students' preferences regarding the curriculum offers, an on line dictionary of general purpose lemma list, an on line terminology glossary, an administrative form related information in GSL. Following the Open eClass pattern, three basic user roles are supported: (i) student, (ii) instructor, and (iii) administrator. However, the platform also supports special intermediary roles such as "administrator assistant", "user administrator", "group leader" and "visitor". These roles serve among other functionalities, the options available for lexical material enrichment through crowd sourcing. The GSL terminology environment allows for the creation of different glossaries directly by their users, where GSL signers are invited to upload their suggestions for various terms under specific quality control conditions. Authorized users may enter new terminology items including the

term definition and various supporting multimedia material (icons, video, text etc), while they can modify or completely delete entries. Furthermore, they can validate terms suggested by non authorized users to make them visible to the whole user community. Terminology enrichment actions incorporate: 1. New lemma or new sense entry 2. Modification of a lemma or a sense 3. Validation of a proposed lemma sense 4. Communication or hiding of a lemma sense or a lemma description 5. Linking of a lemma with a lemma in a different language (Greek and/or English)

Modeling of Geographical Location in French Sign Language from a Semantically Compositional Grammar

Mohamed Nassime Hadjadj

The use of the specificities related to the visuo-gesual modality of SL, such as the use of the signing space and the simultaneous articulation of multiple channels allows the signer to express structures in a more illustrative way. The description of this structure goes beyond the linear linguistic organization initially applied to describe spoken languages. In this paper, we are interested in modeling structures that rely on the signing space to designate the location of one object relative to another. We are particularly interested in the study of location of one place in relation to another one in French Sign Language (LSF). After a presentation of the corpus and the methodology followed to analyze it, we present the study carried out as well as the results obtained.

Raising Awareness for a Korean Sign Language Corpus among the Deaf Community

Sung-Eun Hong, Seongok Won, Il Heo and Hyunhwa Lee

This paper contains strategies that need to be implemented before the sign language community can be involved in corpus work to raise awareness for the need of corpus work. The Korean Sign Language (KSL) Corpus Project began in order to create a linguistic corpus with 60 deaf native and near-native signers form the area of Seoul. In the process of building the KSL Corpus by collecting sign language data and annotating it the project was faced with the challenge that the concept of corpus was completely new to the Korean Deaf community. The KSL Corpus Project developed three strategies in order to inform and explain what the KSL Corpus is about. First, the research project produced numerous KSL videos and posted them on social networking websites in a weekly rhythm. Second, the project organized a workshop, where only deaf people were invited to participate. Third, the KSL Corpus project selected prominent Deaf people who were schooled and provided with corpus materials in order to inform others about KSL Corpus by connecting to their friends and families. The experiences and outcomes of the above strategies are of special importance since the data collection of the KSL Corpus is still in process.

Publishing DGS Corpus Data: Different Formats for Different Needs

Elena Jahn, Reiner Konrad, Gabriele Langer, Sven Wagner and Thomas Hanke

In 2010-2012, the DGS-Korpus project collected a large corpus of German Sign Language (DGS). Now, a substantial subset of the data is published, namely the Public DGS Corpus. We describe the considerations and decisions taken regarding what part of the data is to be made public, the necessary quality assurance measures to the data preparation as well as the formats of the published data. The corpus is published in three different ways in order to fulfil the needs of a variety of different users. First of all, the data is made available to the language community whose members allowed us to share their recorded language. In addition, we hope that a large number of non-scientific users with various backgrounds will find the data useful. Last but not least, we aim to

make the data attractive for users with a scientific background and provide the possibility to conduct studies based on it, irrespective of whether they are familiar with DGS or not.

Where Methods Meet: Combining Corpus Data and Elicitation in Sign Language Research

Vadim Kimmelman, Ulrika Klomp and Marloes Oomen

We discuss three case studies on various grammatical phenomena in Russian Sign Language (RSL) and Sign Language of the Netherlands (NGT) in order to compare corpus-based and elicitation-based approaches to sign linguistics. Firstly, we investigate impersonal reference in RSL using corpus search, informal elicitation, and an acceptability judgment task. Secondly, we examine argument structure and pro-drop licensing in NGT psych verb constructions using corpus search and a supplementary acceptability judgment task. Thirdly, we investigate conditional clauses in NGT based on corpus search, and contrast the findings with those from elicitation-based studies of conditional clauses in other sign languages. The three case studies highlight both the merits and limitations of combining different research methods as well as illustrate some of the issues that arise from doing so – and how they may be navigated. We conclude that corpus-based research serves to identify the boundaries of observed variation and describe both expected and unexpected patterns, while the underlying factors for these patterns can be investigated by eliciting data in more controlled contexts. Finally, we demonstrate that the differences in the results obtained via various research methods have important practical implications, in particular for sign language education.

Workflow Management and Quality Control in the Development of the PJM Corpus: The Use of an Issue-Tracking System

Piotr Mostowski, Anna Kuder, Joanna Filipczak and Paweł Rutkowski

The main goal of the present paper is to describe a workflow management and quality assurance system used in the project of developing the Polish Sign Language (polski język migowy, PJM) Corpus currently underway at the University of Warsaw, Poland. To ensure a satisfactory level of annotation quality, we implemented an external issue tracking system as a basic tool to manage all stages of the annotation process: segmenting the video recording into individual signs, adding glosses to the delineated signs, segmenting text into clauses, translating text into written Polish and adding grammar tags marking different language phenomena. This paper offers a detailed overview of the procedures that we employ, illustrating the most important advantages and disadvantages of our approach and the choices we have made.

Animating AZee Descriptions Using Off-the-Shelf IK Solvers

Fabrizio Nunnari, Michael Filhol and Alexis Heloir

We propose to implement a bottom-up animation solution for the AZee system. No low-level AZee animation system exists yet, which hinders its effective implementation as Sign Language avatar input. This bottom-up approach delivers procedurally computed animations and, because of its procedural nature, it is capable of generating the whole possible range of gestures covered by AZee's symbolic description. The goal is not to compete on the ground of naturalness since movements are bound to look robotic like all bottom-up systems, but its purpose could be to be used as the missing low-level fallback for an existing top-down system. The proposed animation system is built on the top of a freely available 3D authoring tool and takes advantage of the tool's default IK solving routines.

The Cologne Corpus of German Sign Language as L2 (C/CSL2): Current Development Stand

Alejandro Oviedo, Thomas Kaul, Leonid Klinner and Reiner Griebel

Since 2016 (Kaul et al., 2016) a German Sign Language (DGS) learner corpus (Granger et al., 2015) it has been building up at the University of Cologne. Primary data consist of around 60 hours of signed discourse in more than 1,250 individual files produced by 350 DGS hearing learners (312 female / 38 male) whose mother tongue is German. Data has been collected from A1 to C1 CEFR (Council of Europe, 2001) proficiency levels. A similar number of monologues and dialogues is included. Monologues (average duration 2.5 minutes) are mostly induced by an illustration or a video. Dialogues have an average duration of 8 minutes. Dialogues corresponding to the levels A1 to B2 are performed between the informant and a Deaf teacher. At advanced level (C1) dialogues show an interaction between two students. Metadata related to the videos includes age and gender of the informants as well as the proficiency level and semester of data collection. A part of the data corresponds to a longitudinal learner corpus (Granger et al., 2015). This is the case of a group of students who visited DGS-courses of different proficiency levels between mid-2015 and the end of 2017 and were filmed at different times along that period. The corpus is a work in progress. Our primary data are constantly being extended, since each semester new videos are added to the corpus (the tests presented by the students in the DGS courses as well as a number of videos produced and analyzed by the students in linguistics courses). Only around 6% of the videos have received so far transcription: German glosses, translation into German and some linguistic tags have been included in ELAN (Crasborn & Slotjes, 2008) files. Lemmatisation (Johnston, 2010) has been oriented using a lexical database of around 8,000 signs previously produced by our university to serve as teaching material. Current transcriptions also include a series of annotation lines with controlled vocabulary for word-classes, disfluencies (Oviedo et al., in press) and deviations from the DGS standard at phonetic-phonological, morphological and syntactic levels. The biggest challenge faced so far in the development of our corpus is the reluctance of students to authorize the use of the corpus outside our research group. We are only authorized to transcribe the videos and use the transcriptions as a data source. However, a small group of students have up to now authorized us to show their videos and/or video-pictures to external audiences. One strategy that has proved to be useful in obtaining data that can be shared is that of linking students to the tasks of transcription and linguistic analysis. During the 2016/2017 winter semester we held a seminar with masters students to train them in the transcription of their own signed recordings. At the end of the course, the majority of the participants gave us permission to use their videos in public demonstrations.

Crowdsourcing for the Swedish Sign Language Dictionary

Nikolaus Riemer Kankkonen, Thomas Björkstrand, Johanna Mesch and Carl Börstell

In this paper, we describe how we are actively using the Swedish Sign Language (SSL) community in collecting and documenting signs and lexical variation for our language resources, particularly the online Swedish Sign Language Dictionary (SSLD). Apart from using the SSL Corpus as a source of input for new signs and lexical variation in the SSLD, we also involve the community in two ways: first, we interact with SSL signers directly at various venues, collecting signs and judgments about signs; second, we discuss sign usage, lexical variation, and sign formation with SSL signers on social media, particularly through a Facebook group in which we both actively engage in and monitor discussions about SSL. Through these channels, we are able to get direct feedback on our language documentation work and improve on what has become the main lexicographic resource for SSL. We describe the process of simultaneously using corpus data, judgment and elicitation data, and crowdsourcing and discussion groups for enhancing the SSLD, and give examples of findings pertaining to lexical variation resulting from this work.

Improving Lemmatisation Consistency without a Phonological Description. The Danish Sign Language Corpus and Dictionary Project.

Thomas Troelsgård and Jette Kristoffersen

The Danish Sign Language Corpus and Dictionary project at Centre for Sign Language, UCC has a dual aim: to build of Danish Sign Language Corpus, and to use this corpus to expand and improve The Danish Sign Language Dictionary. Our goal is a one-to-one correspondence between sign lemmas in corpus and dictionary, but due to limited resources, we cannot include an accurate phonological description of each sign form. In order to secure a consistent lemmatisation in the corpus as well as across the two resources, we thus rely exclusively on sign videos and Danish equivalents. In this paper, we will describe how we use the lemmas of the Danish Sign Language Dictionary, and additional signs found in connection with the dictionary work, as the initial lexical database of the corpus tool. For new signs found in corpus, the actual corpus tokens will serve as preliminary video representations. To facilitate the sign search when lemmatising corpus tokens, we assign several Danish equivalents to each sign, including all equivalents in the dictionary data. Furthermore, we include synonyms found through linking these equivalents to the Danish wordnet (DanNet), although equivalents added in this way cannot be regarded as valid senses of the sign.

Hand in Hand – Using Data from an Online Survey System to Support Lexicographic Work

Sabrina Wähl, Gabriele Langer and Anke Müller

In the DGS-Korpus project the lexicographic descriptions of signs are based on the available data of the DGS-Korpus, a reference corpus of German Sign Language (DGS). As this corpus is limited in size, number of informants recorded and topics included it is in some cases helpful to obtain additional information from the larger sign language community via an online voting system. This is done using the DGS-Feedback System, a tool especially designed for online surveys conducted using a sign language. With this tool further information on e.g. sign forms and meanings and their use and regional distribution has been elicited. Data from the DGS-Feedback is used in several ways during the lexicographic process of preparing dictionary entries to supplement data from the corpus. In the following the consideration of the data from the DGS-Feedback in relation to the corpus data in decision-making, analysis and lexicographic description is explained and discussed by way of examples.

Exploring Localization for Mouthings in Sign Language Avatars

Rosalee Wolfe, Thomas Hanke, Gabriele Langer, Elena Jahn, Satu Worseck, Julian Bleicken, John McDonald and Sarah Johnson

According to the World Wide Web Consortium (W3C), localization is “the adaptation of a product, application or document content to meet the language, cultural and other requirements of a specific target market”. One requirement necessary for localizing a sign language avatar is creating a capability to produce convincing mouthing. For purposes of this inquiry we make a distinction between mouthings and mouth gesture. The term ‘mouthings’ refers to mouth movements derived from words of a spoken language whereas ‘mouth gesture’ refers to mouth movements not derived from a spoken language. This effort focuses on the former. The prevalence of mouthings varies across different sign languages and individual signers. Although mouthings occur regularly in most sign languages, their significance and status have been a matter of sometimes heated discussions among sign linguists. However, no matter the theoretical viewpoint one takes on the issue of mouthing, one must acknowledge that for most if not all sign languages, mouthings do occur. If an avatar purports to fully and naturally express any sign language, it must have the capacity to express all aspects of the language, which likely will include mouthings. Although most avatar systems

were created for hearing communities, several technologies have emerged to improve speech recognition for those who are hard-of-hearing or who find themselves in noisy environments. These were not satisfactory for Deaf communities as they did not portray sign language. Initial efforts to incorporate mouthing in sign language avatars utilized a mouth picture or viseme for each letter of the International Phonetic Alphabet (IPA), but were hampered by a reliance on blend shapes. Muscle-based avatars have the advantage of avoiding the limitations of blend shapes. This paper reports on a first step to identify the requirements for extending a muscle-based avatar to incorporate mouthings in multiple sign languages.

Session D: Involving the Language Community

Saturday 12 May, 16:30 – 18:00
On Stage Session

Which Picture? A Methodology for the Evaluation of Sign Language Animation Understandability

Vonjiniaina Domohina Malala, Elise Prigent, Annelies Braffort and Bastien Berret

The goal of our study is to explore which information is essential to understand virtual signing. To that aim, we developed an online test to assess the comprehensibility of four different versions of signers: a baseline version with a real human signer, a most complete version of a virtual signer, and two degraded versions of a virtual signer (one with non-visible hands and one without movements of head/trunk). Each video showed the description of a picture in French Sign Language (LSF). After having seen the video, participants had to find which picture had been described among 9 pictures displayed. The originality of our approach was to include two types of confusable pictures on the response board. One was supposed to induce errors by confounding the lexical signs and the other by confounding the spatial structure of the picture. In this way, we explored the effect of hiding hands and blocking trunk/head on the comprehension of lexicon and spatial structure.

The Hong Kong Sign Language Browser

Felix Sze, Kloris Lau and Kevin Yu

This paper describes the design of the Hong Kong Sign Language Browser which was established for providing accessible online resources on the lexical variations of HKSL in order to support the promotion of sign language and other sign-related services in the local community. With continuous funding support from the government since 2012, local Deaf organizations and Deaf signers of diverse backgrounds are invited to contribute their sign language knowledge in the data collection and evaluation process. Each year Deaf informants proficient in HKSL are invited to CSLDS to provide signing data to a pre-defined list of lexical targets. Their signing data are analyzed and variants are identified. These video data are then placed in an online platform for local Deaf organizations for rating and comments, and they can contribute data as well if there are additional variants not yet covered in the initial round of data collection. Once finalized, the lexical variants are placed in the Hong Kong Sign Language Browser for free public access. For each lexical target, each variant is indicated by a different color. Variants that are more commonly used and seen by Deaf organizations are listed first whereas the least common variants are listed last.

SLAAASh and the ASL Deaf Communities (or "so many gifs!")

Julie Hochgesang

The project Sign Language Acquisition, Annotation, Archiving and Sharing (SLAAASh) is a model for working with diverse ASL Deaf communities in all stages of the project. In this presentation, I highlight key steps in achieving this level of collaboration. First, I discuss the importance of sharing work with the community—a key form of reciprocity recognized by Deaf community members. Second, I discuss the importance of reflecting diversity, e.g., ensuring that ASL Signbank actors vary in age, gender, ethnicity, body type, and language experience. Third, I discuss the importance of incorporating feedback from stakeholders and show how the ASL Signbank actors have expressed different views that have impacted our development of the Signbank. Finally, I discuss the crucial component of building substantive community connections and maintaining them long-term. I end by discussing our own efforts to build community connections to date as well as planned future.