7th Workshop on the Representation and Processing of Sign Languages:

Corpus Mining

28 May 2016

ABSTRACTS

Editors:

Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, Johanna Mesch

Workshop Programme

09:00 – 10:30	On-stage Session A: *Corpus Mining*
10:30 – 11:00	Coffee break
11:00 – 13:00	Poster Session B: *Corpora and Mining*
13:00 – 14:00	Lunch break
14:00 – 16:00	Poster Session C: *New Challenges for SL Corpora and Resources*
16:00 – 16:30	Coffee break
16:30 – 18:00	On-stage Session D: *SL Resources: Collaboration and Sharing*

Workshop Organizers

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 Capital, Copenhagen DK
Johanna Mesch	Stockholm University, Stockholm SE

Workshop Programme Committee

Penny Boyes Braem	Center for Sign Language Research, Basel CH
Annelies Braffort	LIMSI/CNRS, Orsay FR
Onno Crasborn	Radboud University, Nijmegen NL
Athanasia-Lida Dimou	Institute for Language and Speech Processing, Athens GR
Sarah Ebling	Institute of Computational Linguistics, University of Zurich, 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
Thomas Hanke	Institute of German Sign Language, University of Hamburg, Hamburg DE
Julie Hochgesang	Gallaudet University, Washington US
Matt Huenerfauth	Rochester Institute of Technology, Rochester, NY, USA
Hernisa Kacorri	City University New York (CUNY), New York, USA
Athanasios Katsamanis	Computer Vision, Speech Communication and Signal Processing Group, National Technical University of Athens, Athens GR
Jette Kristoffersen	Centre for Sign Language, University College Capital, Copenhagen DK
John McDonald	DePaul University, Chicago US
Johanna Mesch	Stockholm University, Stockholm SE
Carol Neidle	Boston University, Boston US
Rosalee Wolfe	DePaul University, Chicago US

Session A: Corpus Mining

Saturday 28 May, 09:00 – 10:30
Chairperson: Julie Hochgesang On-stage Session

The Importance of 3D Motion Trajectories for Computer-based Sign Recognition

Mark Dilsizian, Zhiqiang Tang, Dimitri Metaxas, Matt Huenerfauth and Carol Neidle

Computer-based sign language recognition from video is a challenging problem because of the spatiotemporal complexities inherent in sign production and the variations within and across signers. However, linguistic information can help constrain sign recognition to make it a more feasible classification problem. We have previously explored recognition of linguistically significant 3D hand configurations, as start and end handshapes represent one major component of signs; others include hand orientation, place of articulation in space, and movement. Thus, although recognition of handshapes (on one or both hands) at the start and end of a sign is essential for sign identification, it is not sufficient. Analysis of hand and arm movement trajectories can provide additional information critical for sign identification. In order to test the discriminative potential of the hand motion analysis, we performed sign recognition based exclusively on hand trajectories while holding the handshape constant. To facilitate this evaluation, we captured a collection of videos involving signs with a constant handshape produced by multiple subjects; and we automatically annotated the 3D motion trajectories. 3D hand locations are normalized in accordance with invariant properties of ASL movements. We trained time-series learning-based models for different signs of constant handshape in our dataset using the normalized 3D motion trajectories. Results show significant computer-based sign recognition accuracy across subjects and across a diverse set of signs. Our framework demonstrates the discriminative power and importance of 3D hand motion trajectories for sign recognition, given known handshapes.

Towards a Visual Sign Language Corpus Linguistics

Thomas Hanke

Visualisations have a long tradition in linguistics, as in many fields dealing with complex structure. New forms of representations have been introduced to Visual Linguistics in the recent past, e.g. to help the researcher find the needle in a haystack, i.e. corpus. Here we present visualisation services available in iLex making a combined corpus and lexical database visually accessible. While many approaches suggested for textual languages transfer to sign language data as well, others explore sign-specific structure, such as multi-dimensional concordances not being restricted to sequentiality. Experimental combinations of animated visualisation and image processing might support the researcher to compensate for incomplete high-quality (=manual) annotation. In the long run, we see the potential that visualisation and data manipulation go hand in hand, allowing future user interfaces that are less text-heavy than today's sign language annotation environments.

Using Sign Language Corpora as Bilingual Corpora for Data Mining: Contrastive Linguistics and Computer-assisted Annotation

Laurence Meurant, Anthony Cleve and Onno Crasborn

More and more sign languages nowadays are now documented by large-scale digital corpora. But exploiting sign language (SL) corpus data remains subject to the time consuming and expensive manual task of annotating. In this paper, we present an ongoing research that aims at testing a new approach to better mine SL data. It relies on the methodology of corpus-based contrastive linguistics, exploiting SL corpora as bilingual corpora. We present and illustrate the main improvements we foresee in developing such an approach: downstream, for the benefit of the linguistic description and the bilingual (signed - spoken) competence of teachers, learners and the

users; and upstream, in order to enable the automatisation of the annotation process of sign language data. We also describe the methodology we are using to develop a concordancer able to turn SL corpora into searchable translation corpora, and to derive from it a tool support to annotation.

Session B: Corpora and Mining

Saturday 28 May, 11:00 – 13:00
Chairperson: Johanna Mesch Poster Session

Visualizing Lects in a Sign Language Corpus: Mining Lexical Variation Data in Lects of Swedish Sign Language

Carl Börstell and Robert Östling

In this paper, we discuss the possibilities for mining lexical variation data across (potential) lects in Swedish Sign Language (SSL). The data come from the SSL Corpus (SSLC), a continuously expanding corpus of SSL, its latest release containing 43307 annotated sign tokens, distributed over 42 signers and 75 time-aligned video and annotation files. After extracting the raw data from the SSLC annotation files, we created a database for investigating lexical distribution/variation across three possible lects, by merging the raw data with an external metadata file, containing information about the age, gender, and regional background of each of the 42 signers in the corpus. We go on to present a first version of an easy-to-use graphical user interface (GUI) that can be used as a tool for investigating lexical variation across different lects, and demonstrate a few interesting finds. This tool makes it easier for researchers and non-researchers alike to have the corpus frequencies for individual signs visualized in an instant, and the tool can easily be updated with future expansions of the SSLC.

Linking Lexical and Corpus Data for Sign Languages: NGT Signbank and the Corpus NGT

Onno Crasborn, Richard Bank, Inge Zwitserlood, Els van der Kooij, Anique Schüller, Ellen Ormel, Ellen Nauta, Merel van Zuilen, Frouke van Winsum and Johan Ros

How can lexical resources for sign languages be integrated with corpus annotations? We answer this question by discussing an increasingly frequent scenario for sign language resources, where the lexical data are stored in an online lexical database that may also serve as a sign language dictionary, while the annotation data are offline files in the ELAN Annotation Format (EAF). There is by now broad consensus on the need for ID-glosses in corpus annotation, which in turn requires having at least a list of ID-glosses with a description of the phonological form and meaning of the signs. There is less of a consensus on standards for glossing, on practices of sign lemmatisation, and on the types of information that need to be stored in the lexical database. This paper contributes to the establishment of standards for sign language resources by discussing how two data resources for Sign Language of the Netherlands (NGT) are currently being integrated, using the ELAN annotation software for corpus annotation and an adaptation of the Auslan Signbank software as a lexical database. We discuss some of the present relations between two large NGT data sets, and outline some future developments that are foreseen.

From a Sign Lexical Database to an SL Golden Corpus – the POLYTROPON SL Resource

Eleni Efthimiou, Evita Fotinea, Athanasia - Lida Dimou, Theodore Goulas, Panagiotis Karioris, Kyriaki Vasilaki, Anna Vacalopoulou and Michalis Pissaris

The POLYTROPON lexicon resource is being created in an attempt i) to gather and recapture already available lexical resources of Greek Sign Language (GSL) in an up-to-date homogeneous manner, ii) to enrich these resources with new lemmas, and iii) to end up with a multipurpose-multiuse resource which can be equally exploited in end user oriented educational/communication services and in supporting various SL technologies. The database that hosts the newly acquired resource, incorporates various SL oriented fields of information, including information on compounding, GSL synonyms, classifier qualities, lemma related senses, semantic groupings etc, and also lemma coding for their manual and non-manual articulation activity. It also provides linking of GSL and Modern Greek equivalent(s) lemma pairs to serve bilingual use purposes. A by-product of considerable value is the parallel corpus which derived from the GSL examples of use accompanying each lemma entry in the dictionary and their translations into Modern Greek. The annotation of the corpus for the entailed signs and assignment of respective glosses in combination with data capturing by both HD and Kinect cameras in three repetitions, allowed for the creation of a golden parallel corpus available to the community of SL technologies for experimentation with various approaches to SL recognition, MT and information retrieval.

Annotated Video Corpus on FinSL with Kinect and Computer-vision Data

Tommi Jantunen, Outi Pippuri, Tuija Wainio, Anna Puupponen and Jorma Laaksonen

This paper presents an annotated video corpus of Finnish Sign Language (FinSL) to which has been appended Kinect and computer-vision data. The video material consists of signed retellings of the stories Snowman and Frog, where are you?, elicited from 12 native FinSL signers in a dialogue setting. The recordings were carried out with 6 cameras directed toward the signers from different angles, and 6 signers were also recorded with one Kinect motion and depth sensing input device. All the material has been annotated in ELAN for signs, translations, grammar and prosody. To further facilitate research into FinSL prosody, computer-vision data describing the head movements and the aperture changes of the eyes and mouth of all the signers has been added to the corpus. The total duration of the material is 45 minutes and that part of it that is permitted by research consents is available for research purposes via the LAT online service of the Language Bank of Finland. The paper briefly demonstrates the linguistic use of the corpus.

Methods for Recognizing Interesting Events within Sign Language Motion Capture Data

Pavel Jedlička, Zdeněk Krňoul and Miloš Železný

Rising popularity of motion capture in movie-production makes this technology more robust and more accessible. Utilization of this technology for sign language capturing and analysis is evident. The article deals with the usability of the motion capture in creating sign language corpora. A large amount of the data acquired by the motion capture has to be processed to provide usable data for wide range of research areas: e.g. sign language recognition, translation, synthesis, linguistics, etc. The aim of this article is to explore possible methods to detect interesting events in data using machine learning techniques. The result is a method for detection of the beginning and the end of the sign, hand location, finger and palm orientation, whether the sign is one or two handed, and symmetry in the two-handed signs.

Centroid-Based Exemplar Selection of ASL Non-Manual Expressions using Multidimensional Dynamic Time Warping and MPEG4 Features

Hernisa Kacorri, Ali Raza Syed, Matt Huenerfauth and Carol Neidle

We investigate a method for selecting recordings of human face and head movements from a sign language corpus to serve as a basis for generating animations of novel sentences of American Sign Language (ASL). Drawing from a collection of recordings that have been categorized into various types of non-manual expressions (NMEs), we define a method for selecting an exemplar recording of a given type using a centroid-based selection procedure, using multivariate dynamic time warping (DTW) as the distance function. Through intra- and inter-signer methods of evaluation, we demonstrate the efficacy of this technique, and we note useful potential for the DTW visualizations generated in this study for linguistic researchers collecting and analyzing sign language corpora.

The Usability of the Annotation

Jarkko Keränen, Henna Syrjälä, Juhana Salonen and Ritva Takkinen

Several corpus projects for sign languages have tried to establish conventions and standards for the annotation of signed data. When discussing corpora, it is necessary to develop a way of considering and evaluating holistically the features and problems of annotation. This paper aims to develop a conceptual framework for the evaluation of the usability of annotations. The purpose of the framework is not to give conventions for annotating but to offer tools for the evaluation of the usability of the annotation, in order to make annotations more usable and make it possible to justify and explain decisions about annotation conventions. Based on our experience of annotation in the corpus project of Finland's Sign Languages (CFINSL), we have developed six principles for the evaluation of annotation. In this article, using these six principles, we evaluate the usability of the annotations in CFINSL and other corpus projects. The principles have offered benefits in CFINSL: We are able to evaluate our annotations more systematically and holistically than ever before. Our work can be seen as an effort to bring a framework of usability to corpus work.

Transitivity in RSL: A Corpus-based Account

Vadim Kimmelman

A recent typological study of transitivity Haspelmath (2015) demonstrated that verbs can be ranked according to transitivity prominence, that is, according to how likely they are to be transitive cross-linguistically. This ranking can be argued to be cognitively rooted (based on the properties of the events and their participants) or frequency-related (based on the frequency of different types of events in the real world). Both types of explanation imply that the transitivity ranking should apply across modalities. To test it, we analysed transitivity of frequent verbs in the corpus of Russian Sign Language by calculating the proportion of overt direct and indirect objects and clausal complements. We found that transitivity as expressed by the proportion of overt direct objects is highly positively correlated with the transitive prominence determined cross-linguistically. We thus confirmed the modality-independent nature of transitivity ranking.

Automatic Alignment of HamNoSys Subunits for Continuous Sign Language Recognition

Oscar Koller, Hermann Ney and Richard Bowden

This work presents our recent advances in the field of automatic processing of sign language corpora targeting continuous sign language recognition. We demonstrate how generic annotations at the articulator level, such as HamNoSys, can be exploited to learn subunit classifiers. Specifically, we explore cross-language-subunits of the hand orientation modality, which are trained on isolated signs of publicly available lexicon data sets for Swiss German and Danish sign language and are applied to continuous sign language recognition of the challenging RWTH-PHOENIX-Weather

corpus featuring German sign language. We observe a significant reduction in word error rate using this method.

Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing

Gabriele Langer, Thomas Troelsgård, Jette Kristoffersen, Reiner Konrad, Thomas Hanke and Susanne König

In a combined corpus-dictionary project, you would need one lexical database that could serve as a shared "backbone" for both corpus annotation and dictionary editing, but it is not that easy to define a database structure that applies satisfactorily to both these purposes. In this paper, we will exemplify the problem and present ideas on how to model structures in a lexical database that facilitate corpus annotation as well as dictionary editing. The paper is a joint work between the DGS Corpus Project and the DTS Dictionary Project. The two projects come from opposite sides of the spectrum (one adjusting a lexical database grown from dictionary making for corpus annotating, one building a lexical database in parallel with corpus annotation and editing a corpus-based dictionary), and we will consider requirements and feasible structures for a database that can serve both corpus and dictionary.

A New Tool to Facilitate Prosodic Analysis of Motion Capture Data and a Datadriven Technique for the Improvement of Avatar Motion

John McDonald, Rosalee Wolfe, Ronnie Wilbur, Robyn Moncrief, Evie Malaia, Sayuri Fujimoto, Souad Baowidan and Jessika Stec

Researchers have been investigating the potential rewards of utilizing motion capture for linguistic analysis, but have encountered challenges when processing it. A significant problem is the nature of the data: along with the signal produced by the signer, it also contains noise. The first part of this paper is an exposition on the origins of noise and its relationship to motion capture data of signed utterances. The second part presents a tool, based on established mathematical principles, for removing or isolating noise to facilitate prosodic analysis. This tool yields surprising insights into a data-driven strategy for a parsimonious model of life-like appearance in a sparse key-frame avatar.

The French Belgian Sign Language Corpus. A User-Friendly Corpus Searchable Online

Laurence Meurant, Aurélie Sinte and Eric Bernagou

This paper presents the first large-scale corpus of French Belgian Sign Language (LSFB) available via an open access website (www.corpus-lsfb.be). Visitors can search within the data and the metadata. Various tools allow the users to find sign language video clips by searching through the annotations and the lexical database, and to filter the data by signer, by region, by task or by keyword. The website includes a lexicon linked to an online LSFB dictionary.

Sign Classification in Sign Language Corpora with Deep Neural Networks

Lionel Pigou, Mieke Van Herreweghe and Joni Dambre

Automatic and unconstrained sign language recognition (SLR) in image sequences remains a challenging problem. The variety of signers, backgrounds, sign executions and signer positions makes the development of SLR systems very challenging. Current methods try to alleviate this complexity by extracting engineered features to detect hand shapes, hand trajectories and facial expressions as an intermediate step for SLR. Our goal is to approach SLR based on feature learning rather than feature engineering. We tackle SLR using the recent advances in the domain of deep learning with deep neural networks. The problem is approached by classifying isolated signs from

the Corpus VGT (Flemish Sign Language Corpus) and the Corpus NGT (Dutch Sign Language Corpus). Furthermore, we investigate cross-domain feature learning to boost the performance to cope with the fewer Corpus VGT annotations.

A Digital Moroccan Sign Language STEM Thesaurus

Abdelhadi Soudi and Corinne Vinopol

This paper presents a gesture-based linguistic approach to assisting Moroccan Sign Language (MSL) users in understanding and appropriately using Science, Technology, Engineering and Mathematics (STEM) terminology by creating the first-ever digital MSL STEM Thesaurus. The thesaurus enables Deaf individuals to describe signs and obtain Standard Arabic word equivalents, concept graphics, and definitions in both MSL and Arabic. This is accomplished not only by providing words comparable to signs that they know, but also by providing other information (e.g., signed definitions) that helps differentiate Arabic word choices. The thesaurus is supported by a Concordancer for better illustration and disambiguation of STEM terms. The thesaurus will likely prove to be an invaluable tool that will enable children and adults who rely on MSL for communication, both deaf and otherwise communication impaired, to better understand and write knowledgeably and clearly on STEM topics, and pass standardized assessments.

Online Concordancer for the Slovene Sign Language Corpus SIGNOR

Špela Vintar and Boštjan Jerko

We present the first version of an online concordancing tool for the Slovene Sign Language SIGNOR corpus. The corpus search tool allows querying the SIGNOR annotated database by glosses and displays the hits in a keyword-in-context (KWIC) format, accompanied by frequency information, HamNoSys transcription and metadata. The main purpose of the tool is linguistic research, more specifically sign language lexicography, but also providing general public access to the corpus.

Session C: New Challenges for SL Corpora and Resources

Saturday 28 May, 14:00 – 16:00
Chairperson: Jette Kristoffersen Poster Session

The SIGNificant Chance Project and the Building of the First Hungarian Sign Language Corpus

Csilla Bartha, Margit Holecz and Szabolcs Varjasi

The Act CXXV of 2009 on Hungarian Sign Language and the Use of Hungarian Sign Language recognizes Hungarian Sign Language (HSL) as an independent natural language, moreover it provides the legal framework to introduce bilingual education (HSL-Hungarian) in 2017. In order to establish the linguistic background for bilingual education it was crucial to carry out linguistic research on HSL, which research should be sociolinguistically underpinned and should include corpus-based research. This research also aims to standardize HSL for educational purposes with the highest possible degree of community engagement. During the SIGNificant Chance project a

sign language corpus (approximately 1750 hours) was created. A nation-wide fieldwork was conducted (five regions, nine venues). 147 sociolinguistic interviews and 27 grammatical tests (with 54 participants) were recorded in multiple-camera settings. There were also Hungarian competency tests and narrative interviews conducted with selected participants in order to make the complex description of their different linguistic practices in different discursive contexts possible. We are using ELAN and three different templates to analyze the collected data for different purposes (sociolinguistic-grammatical template, another for short term project purposes, and one for the dictionary). Some parts of the annotation work has been finished which contributed to the writing of the basic grammar of HSL and the creation of a small corpus-based dictionary of HSL.

Collecting and Analysing a Motion-Capture Corpus of French Sign Language

Mohamed-El-Fatah Benchiheub, Bastien Berret and Annelies Braffort

This paper presents a 3D corpus of motion capture data on French Sign Language (LSF), which is the first one available for the scientific community for pluridisciplinary studies. The paper also exhibits the usefulness of performing kinematic analysis on the corpus. The goal of the analysis is to acquire informative and quantitative knowledge for the purpose of better understanding and modelling LSF movements. Several LSF native signers are involved in the project. They were asked to describe 25 pictures in a spontaneous way while the 3D position of various body parts was recorded. Data processing includes identifying the markers, interpolating the information of missing frames, and importing the data to an annotation software to segment and classify the signs. Finally, we present the results of an analysis performed to characterize information-bearing parameters and use them in a data mining and modelling perspective.

Digging into Signs: Emerging Annotation Standards for Sign Language Corpora

Kearsy Cormier, Onno Crasborn and Richard Bank

This paper describes the creation of annotation standards for glossing sign language corpora as part of the Digging into Signs project (2014-2015). This project was based on the annotation of two major sign language corpora, the BSL Corpus (British Sign Language) and the Corpus NGT (Sign Language of the Netherlands). The focus of the gloss annotations in these data sets was in line with the starting point of most sign language corpora: to make general corpus annotation maximally useful regardless of the particular research focus. Therefore, the joint annotation guidelines that were the output of the project focus on basic annotation of hand activity, aiming to ensure that annotations can be made in a consistent way irrespective of the particular sign language. The annotation standard provides annotators with the means to create consistent annotations for various types of signs that in turn will facilitate cross-linguistic research. At the same time, the standard includes alternative strategies for some types of signs. In this paper we outline the key features of the joint annotation conventions arising from this project, describe the arguments around providing alternative strategies in a standard, as well as discuss reliability measures and improvement to annotation tools.

Recognition of Sign Language Hand Shape Primitives With Leap Motion

Burcak Demircioğlu, Güllü Bülbül and Hatice Köse

In this study, a rule based heuristic method is proposed to recognize the primitive hand shapes of Turkish Sign Language (TID) which are sensed by a Leap Motion device. The hand shape data set was also tested with selected machine learning method (Random Forest), and the results of two approaches were compared. The proposed system required less data than the machine learning method, and its success rate was higher.

Linking a Web Lexicon of DSGS Technical Signs to iLex

Sarah Ebling and Penny Boyes Braem

A website for a lexicon of Swiss German Sign Language equivalents of technical terms was developed several years ago using Flash technology. In the intervening years, the backend research database was migrated from FileMaker to iLex. Here, we report on the development of a web platform that provides access to the same technical signs by extracting the relevant information directly from iLex. This new platform has many advantages: New sets of signs for technical terms can be added or existing ones modified in iLex at any time, and changes are reflected in the web platform upon refreshing the browser. Just as importantly, the new platform can now also be accessed through all major mobile operating systems, as it does not rely on Flash. We describe how information on the glosses, keywords, videos of citation forms, status, and uses of the technical signs is represented in iLex and how the corresponding web platform was built.

Juxtaposition as a Form Feature - Syntax Captured and Explained rather than Assumed and Modelled

Michael Filhol and Mohamed Nassime Hadjadj

In this article, we report on a study conducted to further the design a formal grammar model (AZee), confronting it to the traditional notion of syntax along the way. The model was initiated to work as an unambiguous linguistic input for signing avatars, accounting for all simultaneous articulators while doing away with the generally assumed and separate levels of lexicon, syntax, etc. Specifically, the work presented here focused on juxtaposition in signed streams (a fundamental feature of syntax), which we propose to consider as a mere form feature, and use it as the starting point of data-driven searches for grammatical rules. The result is a tremendous progress in coverage of LSF grammar, and fairly strong evidence that our initial goal is attainable. We give concrete examples of rules, and a clear illustration of the recursive mechanics of the grammar producing LSF forms, and conclude with theoretical remarks on the AZee paradigm in terms of syntax, word/sign order and the like.

Examining Variation in the Absence of a 'Main' ASL Corpus: The Case of the Philadelphia Signs Project

Jami N. Fisher, Julie Hochgesang and Meredith Tamminga

The Philadelphia Signs Project emerged from the community's desire to document their local ASL variety, originating at the Pennsylvania School for the Deaf. This variety is anecdotally reported to be notably different from other ASL varieties. This project is founded upon the consistent observations of this marked difference. We aim to uncover what, if anything, makes the Philadelphia variety distinct from other varieties in the United States. Beyond some lexical items, it is unknown what linguistic features mark this variety as "different." Comparison to other ASL varieties is difficult given the absence of a main and representative ASL corpus. This paper describes our sociolinguistic data collection methods, annotation procedures, and archiving approach. We summarize several preliminary observations about potentially dialect-specific features beyond the lexicon, such as unusual phonological alternations and word orders. Finally, we outline our plans to test these features with surveys for non-Philadelphians using Philadelphia lexical items, extending to more abstract phonological and syntactic features. This line of inquiry supplements our current archiving practices, facilitating comparison with a main corpus in the future. We maintain that even without a main corpus for comparison, it is essential to document a language variety when the community wishes to preserve it.

Slicing your SL data into Basic Discourse Units (BDUs). Adapting the BDU model (syntax + prosody) to Signed Discourse

Sílvia Gabarró-López and Laurence Meurant

This paper aims to propose a model for the segmentation of signed discourse by adapting the Basic Discourse Units (BDU) Model. This model was conceived for spoken data and allows the segmentation of both monologues and dialogues. It consists of three steps: delimiting syntactic units on the basis of the Dependency Grammar (DG), delimiting prosodic units on the basis of a set of acoustic cues, and finding the convergence point between syntactic and prosodic units in order to establish BDUs. A corpus containing data from French Belgian Sign Language (LSFB) will be firstly segmented according to the principles of the DG. After establishing a set of visual cues equivalent to the acoustic ones, a prosodic segmentation will be carried out independently. Finally, the convergence points between syntactic and prosodic units will give rise to BDUs. The ultimate goal of adapting the BDU Model to the signed modality is not only to allow the study of the position of discourse markers (DMs) as in the original model, but also to give an answer to a controversial issue in SL research such as the segmentation of SL corpus data, for which a satisfactory solution has not been found so far.

Evaluating User Experience of the Online Dictionary of the Slovenian Sign Language

Ines Kozuh, Primož Kosec and Matjaž Debevc

The extensive use of mobile devices and tablets has resulted in an increasing need for the ubiquitous availability of different types of dictionaries online. The purpose of our study was to evaluate the user experience and usability of the online dictionary of the Slovenian sign language. Six Slovenian hearing non-signers were included in the study. While using the online dictionary, participants were asked to complete six tasks: searching for a letter, a word, written explanation of the word, thematic section and particular fairy tale, as well as completing the quiz. In addition, the participants evaluated the usability of the online dictionary with the System Usability Scale. The findings revealed that participants perceived the tasks "searching for the word" and "searching for the thematic section" to be the most difficult tasks and "completing the quiz" to be the easiest one. Regarding the time measured, the task "searching for the word" was the most time-consuming and the task "searching for the letter" was the least. This study provides insights into how Slovenian hearing users perceive using the online dictionary of the Slovenian sign language and could be the basis for future research with users of Slovenian sign language.

Semiautomatic Data Glove Calibration for Sign Language Corpora Building

Zdeněk Krňoul, Jakub Kanis, Miloš Železný and Luděk Müller

The article deals with a recording procedure for sign language dataset building mainly for avatar synthesis systems. Combined data glove and optical capture technique is considered. We present initial experiences with the motion capture data produced by the CyberGlove3 gloves and a set of new tools to ease the recording process, glove calibration and proper interpretation by the 3D model. It results in a more flexible solution for the sign language capture integrating manual glove calibration with an automatic initialization, time synchronization and high-resolution sensor readings.

"Non-tokens": When Tokens Should not Count as Evidence of Sign Use

Gabriele Langer, Thomas Hanke, Reiner Konrad and Susanne König

Lemmatised corpora consist of tokens as instantiations of signs (types). Tokens usually count as evidences of the signs' use. Frequency of tokens is an important criterion for the lexical status of a

sign. In combination with metadata on the signers' sociolinguistic backgrounds such as age, gender, and origin these tokens can also be analysed for regional and sociolinguistic variation. However, corpora may also contain instances of sign use that do not reflect the sign use of the person uttering them. This is particularly true for metalinguistic discussions of signs, malformed signing and slips of the hand as well as other phenomena such as copying/repeating signs of the interlocutors or from stimulus material. In our presentation we list and discuss different kinds of sign use (tokens) that should either not be counted as proof of a sign type at all or at least not as evidence of regular sign use by that particular person. Examples of these "non-tokens" are either taken from the DGS Corpus or from uploaded video answers of the DGS Feedback. We also discuss some implications on how to annotate these cases.

Creating Corpora of Finland's Sign Languages

Juhana Salonen, Ritva Takkinen, Anna Puupponen, Henri Nieminen and Outi Pippuri

This paper discusses the process of creating corpora of the sign languages used in Finland, Finnish Sign Language (FinSL) and Finland-Swedish Sign Language (FinSSL). It describes the process of getting informants and data, editing and storing the data, the general principles of annotation, and the creation of a web-based lexical database, the FinSL Signbank, developed on the basis of the NGT Signbank, which is a branch of the Auslan Signbank. The corpus project of Finland´s Sign Languages (CFINSL) started in 2014 at the Sign Language Centre of the University of Jyväskylä. Its aim is to collect conversations and narrations from 80 FinSL users and 20 FinSSL users who are living in different parts of Finland. The participants are filmed in signing sessions led by a native signer in the Audio-visual Research Centre at the University of Jyväskylä. The edited material is stored in the IDA storage service produced by the CSC – IT Center for Science, and the metadata will be saved into CMDI metadata. Every informant is asked to sign a consent form where they state for what kinds of purposes their signing can be used. The corpus data are annotated using the ELAN tool. At the moment, annotations are created on the levels of glosses and translation.

Session D: SL Resources: Collaboration and Sharing

Saturday 28 May, 16:30 – 18:00
Chairperson: Thomas Hanke On-stage Session

Towards an Annotation of Syntactic Structure in the Swedish Sign Language Corpus

Carl Börstell, Mats Wiren, Johanna Mesch and Moa Gärdenfors

This paper describes on-going work on extending the annotation of the Swedish Sign Language Corpus (SSLC) with a level of syntactic structure. The basic annotation of SSLC in ELAN consists of six tiers: four for sign glosses (two tiers for each signer; one for each of a signer's hands), and two for written Swedish translations (one for each signer). In an additional step by Östling et al. (2015), all glosses of the corpus have been further annotated for parts of speech. Building on the previous steps, we are now developing annotation of clause structure for the corpus, based on meaning and form. We define a clause as a unit in which a predicate asserts something about one or more elements (the arguments). The predicate can be a (possibly serial) verbal or nominal. In addition to predicates and their arguments, criteria for delineating clauses include non-manual

features such as body posture, head movement and eye gaze. The goal of this work is to arrive at two additional annotation tier types in the SSLC: one in which the sign language texts are segmented into clauses, and the other in which the individual signs are annotated for their argument types.

Preventing Too Many Cooks from Spoiling the Broth: Some Questions and Suggestions for Collaboration between Projects in iLex

Penny Boyes Braem and Sarah Ebling

Collaborative development of sign language resources is fortunately becoming increasingly common. In the spirit of collaboration, having one shared lexicon for sign language projects is a big advantage. However, this poses challenges to aspects pertaining to consistency of data, privacy of informants, and intellectual property. This contribution points out some problems that arise, especially if the common data comes from projects of different institutions. We describe what we have found to be a sustainable legal framework for our collaborative iLex corpus lexicon, giving an overview of the different kinds of partners involved in the creation and exploitation of a shared iLex corpus lexicon and providing our answers to the questions we faced along with an outlook for the future.

Community Input on Re-consenting for Data Sharing

Deborah Chen Pichler, Julie Hochgesang, Doreen Simons and Diane Lillo-Martin

Development of large sign language corpora is on the rise, and online sharing of such corpora promises unprecedented access to high quality sign language data, with significant time-saving benefits for sign language acquisition research. Yet data sharing also brings complex logistical challenges for which few standardized practices exist, particularly with regard to the protection of participant rights. Although some ethical guidelines have been established for large-scale archiving of spoken or transcribed language data, not all of these are feasible for sign language video data, especially given the relatively small and historically vulnerable communities from which sign language data are typically collected. Our primary focus is the process of re-consenting participants whose original informed consent did not address the possibility of sharing their video data. We describe efforts to develop ethically sound, community-supported practices for data sharing and archiving, summarizing feedback collected from two focus groups including a cross-section of community stakeholders. Finally, we discuss general themes that emerged from the focus groups, placing them in the wider context of similar discussions previously published by other researchers grappling with these same issues, with the goal of contributing to best-practices guidelines for data archiving and sharing in the sign language research community.