Hamburg Summer School on Language Documentation and Corpus Technology

Welcome to the programme of the 2019 Hamburg Summer School on Language Documentation and Corpus Technology! This one-week summer school is jointly organized by the Academy of Sciences and Humanities long-term projects DGS-Korpus and INEL in cooperation with the EU-funded The Sign-Hub project.

The summer school starts on Monday, Sept 30, 2019 and ends on Friday, Oct 4, 2019, including Oct 3 which is a public holiday in Germany. The location is the Institute of German Sign Language and Communication of the Deaf at Gorch-Fock-Wall 7, just a couple of minutes walking distance from the University main building.

Each day of the summer school has a specific thematic focus and will start with the morning session giving an introduction to the focus topic covering both signed and spoken language perspectives whereas the afternoon session will deal with modality-specific aspects.
In addition, there are evening lectures on Tuesday and Thursday as well as a social event on Wednesday. For the details, please have a look at the schedule on page 2 as well as the course abstracts on the subsequent pages.

Please register to the summer school via the online form at http://registration.dgs-korpus.de. Registration will remain open until all seats are filled. Registration is free.

On site, the registration desk will be open on Sunday, Sept 29 from 9:00 to 17:30, on Monday from 8:30 to 17:30, and all following days from 9:00 to 17:30.

If you stay on from the TISLR conference ending on Saturday, please note that there are related events on Sunday, Sept 29, also at Gorch-Fock-Wall 7:
09:00–11:00 Presentation of Release 2 of the Public DGS Corpus (with breakfast)
11:00–18:00 Workshop “Sign Language Translation and Avatar Technology”
For more information on these events, cf. http://release2.dgs-korpus.de and http://sltat.dgs-korpus.de.

Schedule

	MORNING 09:00–10:30 & 11:00-12:30	AFTERNOON 14:00–15:30 & 16:00–17:30	EVENING 19:00–20:30
MON	Registration (08:30-09:30) ■■ Introduction: Corpus Work in DGS-Korpus and INEL ■■	Describing Sign Language Grammars: Sequential vs. Simultaneous Compounds ■ Describing Spoken Language: From Corpus Data to Grammar ■	
TUE	Lexicography: Researching, Documenting, and Describing the Lexis of Spoken and Signed Languages ■■	Sign Language Lexicography: Corpus-based Dictionary Writing ■ Spoken Language Lexicography: Working from a Speech Corpus ■	Evening Lecture: Fundamental multimodality or language use and its implications for data collection and corpus creation ■■
WED	Corpus Anonymisation: Making your Corpus Ready for Publication ■■	Data Life Cycle: Keeping your Corpus Alive ■■	Social Event ■■
THU	Quality Assurance: Lemma Revision and Annotation Consistency ■■	Tutorial: OpenPose for Linguists ■■ Spoken Language Corpora: Methodology and Practice I ■	Evening Lecture: Avatar Technology: Is it relevant to sign language linguistics? ■■
FRI	Data Visualisation: Using Graphs and more to Evaluate and Present your Data ■■	Tutorial: HamNoSys ■ Spoken Language Corpora: Methodology and Practice II ■	

Please note that the afternoon courses run in parallel.

■ Recommended for participants mainly interested in sign language data, International Sign interpreting provided

■ Recommended for participants mainly interested in spoken language data

Course Abstracts

Introduction: Corpus Work in DGS-Korpus and INEL

The DGS-Korpus and INEL teams

In order to make it easier for participants to get into the content of the summer school and to provide every participant with the same background knowledge, we will briefly describe the DGS-Korpus project as well as the INEL project, including former project phases and plans for the future.
For DGS-Korpus, we will address data collection, a brief introduction into the annotation tool (iLex) and notation system (HamNoSys) used, the annotation workflow and current statistics, as well as other topics regarding the project.
For INEL, we will introduce the principles of data collection, provide a brief introduction to the annotation tools (i. e. FLEx and EXMARaLDA), and present, among other, the workflows used in the project.

Describing Sign Language Grammars: Sequential vs. Simultaneous Compounds

Mirko Santoro (SignHub)

Compounding is a word/sign formation process based on the combination of lexical elements (words/signs or stems/roots) and it has been documented as a key strategy to enrich the lexicon, even in situations of emergent languages. Across the world's spoken languages, these processes tend to be formed via concatenative morphology in the overwhelming majority of cases. However, once one starts looking at the realm of sign language, an immediate contrast arises. Most morphological information and processes are expressed simultaneously with the stem/root/lexeme during the articulation of a sign. I offer a refined and more comprehensive typology of compounds, in which classifiers and simultaneous forms are also taken into account. Building on a model of the ASL lexicon (Brentari & Padden 2001), I provide a formal account that derives the whole typology of compounds found in LIS and LSF. The class also has a practical part in which students will be invited to describe compounds using the guidelines suggested in the SignGram Blueprint: A guide to sign language grammar writing. Berlin: Mouton De Gruyter, 2017 (open access at https://www.degruyter.com/view/product/467598).

Describing Spoken Language: From Corpus Data to Grammar

Eugénie Stapert (Leiden), Chris Lasse Däbritz (INEL)

Grammar writing is considered an integral part of language documentation. During this workshop, participants will be introduced to the basics of grammar writing, using data from spoken language corpora, in particular from the Dolgan corpus of the INEL project. After a general introduction to the field of grammar writing, the Dolgan language and its speakers, the participants will be challenged to solve linguistic problems with the help of hands-on exercises. By doing so they will acquire insight into the procedures and challenges of grammar writing, as well as into the benefits of using corpus data for this purpose.

Lexicography: Researching, Documenting, and Describing the Lexis of Spoken and Signed Languages

Timofey Arkhangelskiy (INEL), Gabriele Langer (DGS-Korpus), Anke Müller (DGS-Korpus), Thomas Schmidt (Mannheim), Sabrina Wähl (DGS-Korpus)

In this course we introduce and compare the documentation and lexicographic description of the words used in spoken (as opposed to written) vocal languages and the signs of signed languages. Where do lexicographers face common problems in spoken language and sign language lexicography and where do they face modality-specific challenges? This course brings together the perspectives of researching the spoken variety of a well-documented written vocal language (e.g. spoken German), the documentation of not well-researched minority or lesser researched languages with or without a written tradition and the lexicographic description of a signed language (e.g. DGS – German Sign Language).

Sign Language Lexicography: Corpus-based Dictionary Writing

Gabriele Langer (DGS-Korpus), Anke Müller (DGS-Korpus), Sabrina Wähl (DGS-Korpus)

In this course we look at the possibilities, advantages and challenges of working with a sign corpus in dictionary-making. We will share our experience in how we use the DGS Corpus at different stages of dictionary preparation, that is e.g. lemmatisation, lemma selection, sign sense discrimination, analysis of usage including regionality, and collocations.
In the second part of this course, a hands-on session, we focus on selected aspects of sign language lexicography, e.g. entry structures (microstructure), the representation of signed content, and the selection and preparation of authentic example sentences.

Spoken Language Lexicography: Working from a Speech Corpus

Timofey Arkhangelskiy (INEL)

With the help of digital technology, corpus based lexicography has become an important way to document and structure the lexicon of endangered languages.
In this workshop, participants will be introduced to the main principles of lexicography using spoken language data gained from fieldwork. After this, they will be trained to put this knowledge into practise, using data from the Beserman corpus (http://beserman.ru/corpus/search/).

Corpus Anonymisation: Making your Corpus Ready for Publication

Elena Jahn (DGS-Korpus), Reiner Konrad (DGS-Korpus), Timm Lehmberg (INEL)

Building a balanced corpus of a signed or spoken language not only requires to choose data from a diverse and widespread spectrum of text types and genres (like in the case of written corpora) but also to query information on personal data like age, gender, biographical data and regional information.
Apart from informed consent, ethical issues arise when this information is made available to third parties. Even in cases where informants agreed on the use of their personal data, any information gained from them has to be treated as sensitive (e.g. when deaf informants report real life experiences) and – depending on usage scenarios – anonymized.

This course will demonstrate the steps necessary to make corpus data ready for publication. We will highlight three aspects: a) the selection process, b) check routines and tests to remove errors and make the data more consistent, and c) anonymisation, with a particular focus on the last point. During the course, there will be lecture-sessions as well as discussions and hands-on sessions.
Furthermore short insights into applicable law will be given in order to enable participants to distinguish between legal (mostly data protection and copyright related) and ethical issues.

Data Life Cycle: Keeping your Corpus Alive

Thomas Hanke (DGS-Korpus), Elena Jahn (DGS-Korpus), Reiner Konrad (DGS-Korpus), Timm Lehmberg (INEL)

To create a corpus that can be used sustainably by a broad spectrum of scientists, the mere publication of the resource in digitized form is not sufficient at all. Several aspects of publicity, usability, visualization and community interaction need to be considered to grant for its successful long-term publication and usage.
A variety of interfaces for accessing and searching the corpus, like for instance graphic user interfaces designed both for experts and laymen, download and visualization functionalities as well as APIs for direct query purposes can contribute to meet these requirements.
In our workshop we will discuss these and other aspects of data sustainability from the angle of the principles defined by the “Data Lifecyle” illustrated by concrete examples from the DGS-Korpus and the INEL project.

Quality Assurance: Lemma Revision and Annotation Consistency

Alexandre Akhipov (INEL), Daniel Jettka (INEL), Reiner Konrad (DGS-Korpus)

The description of recorded language, for instance via initial token-type matching (lemma-tisation) of signed utterances or the transcription/annotation of speech, is an interpretative process which requires checking routines and control steps in order to reach a high level of consistency and reliability.
For most sign languages, token-type matching (lemmatisation) and building a lexical data-base has to go hand in hand. After the first annotation pass, several checks and control steps are needed to assure a high quality of annotation. These steps include correction of mis-matches, completeness of annotation as well as revision of under- and over-differentiation of types, or redefining citation forms.
In spoken language corpora of under-resourced languages, interlinear morpheme glossing presumably is the most common annotation type. Depending on the annotation workflow, the risk of introducing inconsistencies decreases with the degree of automation. Strict version control, controlled or lexicon-based annotation vocabularies, and automatic error-checking routines help to achieve better consistency.
After talks and discussions covering various aspects of corpus building for spoken and sign language, practical exercises will give insight into diverse revision processes.

Tutorial: OpenPose for Linguists

Maren Brumm (DGS-Korpus), Thomas Hanke (DGS-Korpus), Marc Schulder (DGS-Korpus)

When working with multimodal data, most annotation steps are done manually. OpenPose from Carnegie Mellon University opens the door to analyzing the visual domain without

becoming an expert in Computer Vision: In essence, OpenPose determines the position of joints and returns them as time-series data. This tutorial shows you how to work with OpenPose in order to detect handedness, headshake & more and how to feed the results into your annotation, independent of whether you work with sign or gesture.

Spoken Language Corpora: Methodology and Practice I & II ■

The INEL team

This two-part workshop aims at providing basic skills in the creation and usage of spoken language resources as empirical base for the validation of linguistic hypotheses. It will cover a comprehensive spectrum of technological and methodological aspects regarding the creation (including compilation, transcription and annotation) and exploitation of spoken language corpora. The content will therefore range from general discussion of corpus methodology to thematic hands-on lessons.
In an initial part of the workshop, basic terms and concepts of the qualitative and quantitative analysis of language corpora will be outlined. It is intended that – based on this – participants apply these impulses according to their individual research issues. If desired, in-depth introduction to more specific methods and tools will be given at any time.

Data Visualisation: Using Graphs and more to Evaluate and Present your Data ■■

Thomas Hanke (DGS-Korpus), Daniel Jettka (INEL), Marc Schulder (DGS-Korpus)

When analysing a corpus, a great aid to researchers is the ability to visualise data. What means of visualisation exist and when should they be used? How can specific visualisations be created? Can data representations be misleading?
In this course we will discuss these questions. We will begin with basic visualisations such as pie charts, line graphs and histograms, discussing their potentials and pitfalls. As the course progresses, we will move on to more complex visualisation techniques, for example how to use geographic metadata to represent regional language variations on a coloured country map. All these topics will be addressed in interactive exercises in which participants will learn how to create their own visualisations.

Tutorial: HamNoSys ■(■)

Dolly Blanck (DGS-Korpus), Thomas Hanke (DGS-Korpus)

HamNoSys is a phonetic transcription system that should be able to describe the manual components (and a bit more) of any sign language as well as gesture. It plays a central role in the DGS-Korpus project, e.g. for searching signs in the lexical database.
This course teaches you how to read and write HamNoSys as well as how to use the tools available to “type” HamNoSys and to verify it by sending it to an avatar system capable of animating signs or gestures described in HamNoSys.

Evening Lectures

Tue: Fundamental multimodality or language use and its implications for data collection and corpus creation

Mandana Seyfiddinipur

The study of language began with the study of ancient written text. This practice has shaped and is still shaping our practices in language documentation. The way we conceptualise language and language use has implications for our documentary practices. We still focus on what we think can be written down and often disregard what we think cannot be written down.
But, typically, when we speak, we cannot only hear each other but also see each other. Language is grounded in face-to-face interaction and speaking is a joint activity (Clark 1996). Language acquisition is a process that takes place in face-to-face contexts and our cognitive system automatically integrates both what we hear and what we see (McGurk & McDonald 1976). When we speak, we use our hands to gesture and the information provided in this visual, gestural modality is also integrated automatically in our mind. The gestures we use contribute crucially to our understanding of what speakers are communicating (Kendon 2004). Communities have developed alternate sign languages used in e.g. mourning practices (Kendon). Deaf people develop fully fledged sign languages in the manual modality (Meir et al. 2012).
However, despite this basic multimodal nature of language use we often still do not document language to its full extent due to restricting our recordings to audio or restricting video recording to a few genres like story telling. In this talk I will exemplify the multimodal nature of language use, focusing on manual gesture in its various forms and functions from indexing to semantic specification, and discourse structure marking.

I will discuss its implications for language documentation practices. The role of video recording and the way language use needs to be video recorded to provide useable material for linguistic and ethnographic documentation and analysis will be highlighted. A methodology for training the much needed video recording will be suggested which embeds the technical training of video technology and recording within a theoretically grounded understanding of language use.

Wed: Social Event

Thu: Avatar Technology: Is it relevant to sign language linguistics?

Rosalee Wolfe (Chicago), John McDonald (Chicago)

Avatars becoming ever more prevalent in our lives, do they have a role to play in sign linguistics research? We've seen impressive avatars in movies, on the Internet, and in computer games. So why is it so hard to make an avatar that can produce utterances that are natural looking and grammatical?
We explore resources, alternatives, limitations and opportunities for using avatars to produce signed language, including demonstrations of the current state of the art and an invitation to join the collaboration.