

# CLARA Summer School 2010

## Advanced Language Resource Creation, Archiving and Usage

	Monday, 5.7	Tuesday, 6.7	Wednesday, 7.7	Thursday, 8.7	Friday, 9.7	Saturday, 10.7.	Sunday, 11.7
Major Topic	Overview	Research	AV Matters & Creation	AV Analysis	Annotation / Analysis and Enrichments	Summary& Excursion	Summary& Excursion
9.00 - 12.00	Introduction & Trends	Corpora for and in Multimodal & Multilingualism Research	Audio/Video Recording and Encoding Principles	Speech Analysis with advanced tools such as Praat	Annotation Schemes	Working Group Reports and Discussion	Free
Lunch							
14.00 - 17.00	Gesture research, Sign Language	Working Groups Introduction	AV Conversion Tools and Conversions	Advanced Video Analysis Methods and Tools	Web-Presentation Techniques such as GIS Overlays, Portals, Annex etc.	Barbecue	Free
17.00 - 18.00		Working Groups	Working Groups	Working Groups	Working Groups		

	Monday, 12.7	Tuesday, 13.7	Wednesday, 14.7	Thursday, 15.7	Friday, 16.7	
Major Topic	Annotation / Analysis and Enrichments	Standards & Presentation	Uploading & Organizing	Archiving & Curation	Summary and Finish	
9.00 - 12.00	Advanced annotation Schemas and Tools	Standards for the Language Resource Domain	Corpus organizations techniques and metadata descriptions	Long term preservation of language resources, authenticity issues	Working Group Reports and final discussion	
Lunch						
14.00 - 17.00	Advanced corpus analysis and statistics methods and tools	Enrichment techniques and tools to create lexica and conceptual spaces	Resource uploading and IPR, Access permission issues	Quality issues, quality assessment, archiving and curation costs		
17.00 - 18.00	Working Groups	Working Groups	Working Groups	Working Groups		

Most of the sessions include tutorial parts and practical exercises. Each session will raise a number of issues and questions that will be addressed by working groups of participants. These will be discussed during the courses and at the end of each week. The working groups will cover about 5 participants. Each WG should summarize the activities, make comments and raise questions.

Last update: 28/06/2010

**Monday, July 5, 9-12**

**1. Introduction and Trends**

**Peter Wittenburg, Paul Trilsbeek and Alexander König (Max Planck Institute for Psycholinguistics)**

This session will give an overview about the major trends in „Advanced Language Resource Creation, Archiving and Usage” and thus an overview about what will be dealt with in this summer school. The enormous innovation rate in technological development changed the opportunities to create new multimedia or time series recordings, to annotate and to enrich and analyze them in many ways. New software tools are available that allow users to carry out work on notebooks which a decade ago was not possible. As a consequence the challenges of proper data management are a hot issue due to the huge number of data resources people are creating. This session will touch all aspects briefly and indicate what the worldwide trends currently are.

**Monday, July 5, 13-16**

**2a. Gesture Research (Hedda Lausberg)**

In the first part of the workshop, I will introduce some topics in current gesture research. While the question why people gesture, when they talk has been subject to long-standing controversial discussions, neuroscientific research now provides evidence that gesture production is linked to distinct cognitive processes. Specific types of co-speech gestures are related to specialized left and right hemisphere functions, such as rhythm, prosody, symbolic thinking, spatial cognition, or emotional functions. As such, gesture not only reflects, but possibly also contributes to constitute cognitive and emotional processes.

In the second part of the workshop, I will focus on gesture research methodology. Specifically, I will introduce the NEUROGES-ELAN System, which is a behaviour gesture coding system combined with an annotation tool. The NEUROGES coding system consists of three modules which progress from gesture kinetics to gesture function. Grounded on empirical neuropsychological and psychological studies, the theoretical assumption

behind NEUROGES is that its main kinetic and functional movement categories are differentially associated with specific cognitive, emotional, and interactive functions. ELAN is a free, multimodal annotation tool for digital audio and video media (see Friday session on ELAN and TROVA).

**Monday, July 5, 16-17:30**

**Sign Language**

**2b. Onno Crasborn [Centre for Language Studies, Radboud University Nijmegen]**

Signed languages are used in deaf communities throughout the world, but they have only received linguistic attention from researchers in the last fifty years. They will be briefly characterised and contrasted to spoken languages and gesture. Some factors of all signed languages that have an impact on language documentation are the lack of widely used writing systems or even transcription systems, the multitude of manual and non-manual articulators that are involved, the short history of most languages, and the fact that linguistic and gestural elements are expressed through the same articulators. Common methodologies for data collection, transcription, annotation and archiving will be discussed, in close connection to the tools offered by MPI (ARBIL, ELAN, ANNEX, LEXUS).

---

---

**Tuesday, July 6, 9-12 MM & ML Research**

**3a. The annotation and use of multimodal behaviours in human-human conversations**

**Costanza Navarretta (University of Copenhagen)**

I will present the MUMIN annotation scheme for annotating specific communicative functions of non-verbal behaviour such as feedback, turn-taking and sequencing, and present examples of the application of the scheme to multimodal corpora in various languages. Then I will show how the annotated corpora have been used as training data for machine learning

experiments with the aim of identifying specific communicative functions automatically or discovering dependencies between the various types of annotation.

### **3b. Multimodal communication: What do gestures reveal about language, culture and cognition ?**

**Asli Ozyurek (Radboud University/MPI)**

It is a common practice in language studies to focus on speech only to archive and describe new languages or to conduct cross-linguistic comparisons. In this talk I will outline research which shows that gestures that speakers use give additional insights about their language and culture-specific cognition patterns as well as about how they use both modalities (i.e., speech-gesture ensembles) in language-specific ways to express composite messages (Clark, 1996). Thus I will emphasize the need to collect and analyze multimodal data in order to understand the diversity of human communication patterns and the underlying cognitive processes that might give rise to this diversity.

**Tuesday, July 6, 14-17**

### **4. Working Groups introduction**

**Peter Wittenburg, Paul Trilsbeek, Alexander König (MPI)**

We will build three working groups that have the task to wrap up the highlights of the presentations and discussions, identify gaps and badly explained topics and gather open questions of the participants. Each of the group will create a one page report each day which will be taken up at the following day by the mentors of each of the group. The mentors will check how we can best react on the issues mentioned. For reacting on larger issues we may use the slots at the end of each week or change the schedule.

---

---

**Wednesday, July 7, 9-12 and 14-17**

### **5. Audio/Video Recording and Encoding Principles**

**AV Conversion Tools and Conversions**

**Paul Trilsbeek, Florian Wittenburg (MPI)**

This session will present modern audio/video recording, digitization, encoding and conversion techniques. AV Recordings are normally at the beginning of the data life cycle and the source of all research work. Thus they need to be created with care. For researchers it is important to know the different AV formats and encoding standards, since they will determine the resolution and quality of the recordings to a large extent. Since there are so many standards and compression methods it is also important to understand the transformation opportunities and tools that can be used.

---

---

**Thursday, July 8, 9-12**

### **6. Speech Analysis with Praat**

**Paul Boersma (University of Amsterdam)**

Speech signals are important for any understanding of the nature of human communication. Thus it is important for researchers to know about the articulators and their relations to the acoustic events they produce, so that by measuring the parameters of pitch and formants and so on we achieve the goal of being able to say things about what the articulators are doing. PRAAT is currently the most widely used tool to carry out speech analysis. In these three hours, therefore, the whole of speech science will be discussed, including the ways of measuring all that with Praat.

**Thursday, July 8, 14-17**

**7. Video Analysis Methods and Tools**

**Stefano Masneri (Fraunhofer Institut für Nachrichtentechnik, Berlin)**

Sound is one channel of human communication. But there are also the non-verbal communication channels that in most cases are perceived with our eyes. The complementary technologies to microphones are video cameras and specific methods to record time series of data. This session will present the basics of video analysis methods and tools for example to automatically detect gesture behavior. The area of video analysis is a very dynamic one thus the talk will present various work in progress.

---

---

**Friday, July 9, 9-12**

**8. Annotation Schemes, Annotation Tool (ELAN) and Analysis (TROVA)**

**Han Sloetjes (MPI)**

While AV recordings, sometimes in combination with time series (motion capture) data, are the source material for research, annotations are the basis for most analysis and theorizing work. Automatic recognition techniques often cannot be applied successfully to AV recordings and here manually created annotations are indispensable. ELAN is a modern annotation tool that allows users to create multilayer annotations on AV and time series streams of all kinds of communicative modalities. With TROVA researchers can query and analyze these layers of annotations. ELAN allows users to link to entries in concept registries like ISOcat (see Standards for the LR domain) in the process of annotation.

**Friday, July 9, 14-17**

**9. Web-Presentation Techniques such as GIS Overlays, Portals, Annex etc.**

**Paul Trilsbeek (MPI)**

The internet is taking an ever increasing role in the presentation of research findings and in presenting yourself as an academic professional. In this tutorial we will look at various presentation channels such as web sites, blogs, geographical information systems like Google Earth/Maps, online video presentation sites such as YouTube and an online presentation tool for annotated audiovisual material called ANNEX. We will also see how these various channels can be combined in one single web site or portal.

---

---

**Monday, July 12, 9-12**

## **10. Annotation Schemas and Tools**

**Thomas Schmidt (University of Hamburg)**

While the Friday session 9 started with rather practical annotation methods and concrete tools, this session will describe the principles behind annotations and different views on this. It will bring an overview about multimodal annotation structures and give an overview about different requirements and annotation tools. Users should know that there is a variety of good tools out there often optimized towards a certain application domain. Annotations structures once specified by XML schemas can be transformed to a variety of presentation forms.

**Monday, July 12, 14-17**

## **11. Advanced corpus analysis and statistics methods and tools**

**Taras Zakharko (University of Leipzig)**

This session will present the data manipulation tool R which offers a very powerful framework for research work. Being a full-fledged programming language, usage of R requires more insight into technical aspects, such as basic programming knowledge. In return, it's flexibility offers the user virtually unlimited amount of possibilities and allows to perform analysis tasks of arbitrary complexity.

---

---

**Tuesday, July 13, 9-12**

## **12. Standards for the Language Resource Domain**

**Andreas Witt (Institut für Deutsche Sprache, Mannheim), Menzo Windhouwer (MPI)**

Increasingly important for proper lifecycle management and persistent access to resources is the adherence to international standards. For

character encoding UNICODE has made its way, for declaring semantics the new ISOcat system based on the ISO 12620 metamodel seems to become accepted and for a variety of linguistic data types such as lexica and annotations flexible frameworks are currently being standardized. ISOcat is a tool that allows users to create and use concept entries in the annotation process. The hope is that if everyone uses the same registered concepts we can achieve a higher degree of semantic interoperability. Researchers don't have to understand all details, but they should be aware of what currently is being worked out.

**Tuesday, July 13, 14-17**

## **13. Enrichment techniques and tools to create lexica and conceptual spaces**

**Jacquelijnn Ringersma, Huib Verweij, and Andre Moreira (MPI)**

While annotations describe the flow of words, idioms, expressions along a time axis embedded in a linguistic system, lexica present these units isolated and describe their characteristics along several linguistic dimensions. LEXUS is a modern lexicon tool allowing every researcher to encode the information he/she is interested in and to add multimedia fragments such as images, sound or video clips. Conceptual spaces allow people to create a graphical domain where lexicalized items are related with each other and to navigate in such semantic domains and to link from all concepts to all kinds of resource fragments in the archive. VICOS is a tool that allows doing this.

---

---

**Wednesday, July 14, 9-12 and 14-17**

## **14. Corpus organizations techniques and metadata descriptions**

**Dieter van Uytvanck, Peter Withers, Mariano Gardelini (MPI)**

As already indicated we are confronted with an enormous increase of language resources and collections created by many researchers. If we want to ensure that these resources can be found again and that they can be re-

used we need to describe them by metadata and integrate them into a meaningful organization scheme. PIDs need to be associated with resources to give them a unique identity and to enable stable references. Rights need to be associated with each resources and finally they need to be uploaded into a stable repository system. At the example of the MPI archive it will be demonstrated how a proper setup can look like.

---

---

**Thursday, July 15, 9-12**

**15. Long term preservation of language resources, authenticity issues**

**Peter Wittenburg (MPI)**

For many resources long-term preservation is an issue which can mean storage of data for 10 years or for many language resources storage for infinity. Is this something that can be achieved in the digital area? What has been changed and what measures need to be taken to ensure long life time of data. What kind of curation measures need to be taken to keep data interpretable. What are the costs for long-term archiving and what are the IPR issues associated with resources. This session will touch these issues.

**Thursday, July 15, 14-17**

**16. Quality issues, quality assessment, archiving and curation costs**

**Dirk Roorda**

Repositories make statements about their policies, but how can we be sure that they really do what they claim and how can we ensure a high quality of services and thus establish trust at the side of depositors and users. This session will discuss quality issues and present the details of the Data Seal of Approval method to assess in how far quality criteria are met.

---

---