

Annotation Tools and Standards for Diachronic and Multilingual Research

(Prof. Dr. Gabriele Diewald, Timm Lehmborg)

Empirical studies of grammar often require comprehensive resources of digitized spoken and/or written language data (corpora) that are compiled with respect to their particular research question. In most cases, these corpora are enriched using research relevant metadata as well as linguistic annotation on different layers (morphology, syntax, discourse etc.). In recent years, the number of richly annotated specialized corpora to be used for linguistic purposes has increased dramatically. Due to the fact that the compilation of language corpora requires an immense amount of effort, the creators of corpora are well advised to use widely adopted encoding and annotation standards, to keep their research data (as well as the results that are based on them) as reliable and sustainable as possible. An important task that is tied intrinsically to this issue is the development of user-friendly tools and query interfaces that simplify the annotation and querying of the data according to a certain standard. In our workshop we will focus on the application of widely adopted standards for the encoding and annotation of linguistic data like TEI¹ and EXMARaLDA² as well as software tools and web-based platforms that have been developed in order to meet the special demands of diachronic and contrastive language studies. For this purpose, we will give concrete examples coming from the empirical work on a number of specialized corpora whereas each of the oral presentations will focus on a special issue of diachronic and/or multilingual corpora.

Workshop Schedule:

Part I chair: Timm Lehmborg	
9:00	Introduction
9:30	Thomas Schmidt Querying Spoken Language Corpora
10:00	Kai Wörner Annotating Spoken Language Corpora
10:30	Coffee Break
Part II chair: Dr. Thomas Schmidt	
11:00	Gabriele Diewald, Timm Lehmborg Modeling Diachronic Semantic Change
11:30	Peter M. Fischer Aligning Textual Material - Exemplified on Parallel Corpora
12:00	Timm Lehmborg, Georg Rehm, Andreas Witt Sustainability of Richly Annotated Linguistic Corpora

¹ <http://www.tei-c.org>

² <http://www.exmaralda.org>

Querying Spoken Language Corpora

Dr. Thomas Schmidt (SFB 538 / University of Hamburg)
Max-Brauer-Allee 60, D-22765 Hamburg
tel: +49 40 42838-6425
fax: +49 40 42838-6116
email: thomas.schmidt@uni-hamburg.de
keywords: Spoken language, Corpora, Discourse Analysis
language: English
presenter: Dr Thomas Schmidt

Thomas Schmidt's talk will give a demonstration of the software tool EXAKT which takes into account the special requirements spoken language data poses on corpus query. Several corpora of multilingual spoken discourse will be used to illustrate the methodological and technical issues arising in this area.

Annotating Spoken Language Corpora

Kai Wörner, M. A. (SFB 538 / University of Hamburg)
Max-Brauer-Allee 60, D-22765 Hamburg
tel: +49 40 42838-6425
fax: +49 40 42838-6116
email: kw@exmaralda.org
keywords: annotation, speech, transcription
language: English
presenter: Kai Wörner

Kai Wörner's presentation will list some of the problems that hamper the use of existing annotation tools with spoken language corpora. Based on this, a practical approach to deal with these problems when designing an annotation tool for spoken language data will be presented.

Modelling Diachronic Semantic Change

Prof. Dr. Gabriele Diewald (University of Hannover)
Timm Lehmborg, M. A. (University of Hannover)
Königsworther Platz 1, D-30167 Hannover
tel: +49 511 762-19379
fax: +49 511 762-4814
email: gabriele.diewald@germanistik.uni-hannover.de
keywords: grammaticalization, semantic change, diachronic corpora
language: English
presenter: Timm Lehmborg

This presentation will approach the issue of annotating and querying a comprehensive webbased diachronic corpus of German (KALI) that is used for the purposes of university didactics as well as grammaticalization research. After presenting the principles of annotation and lemmatization of the diachronic data, some proposals on modelling diachronic semantic change will be discussed.

Aligning Textual Material - Exemplified on Parallel Corpora

Dipl. Inf. Peter M. Fischer (SFB 538 / University of Hamburg)

Max-Brauer-Allee 60, D-22765 Hamburg

tel: +49 40 42838-6425

fax: +49 40 42838-6116

email: peter.m.fischer@uni-hamburg.de

keywords: multilingualism, parallel corpora, TEI

language: English

presenter: Peter M. Fischer

Peter M. Fischer will talk about some issues concerned with the alignment of textual material. Using the example of parallel corpora, related possibilities offered by the TEI standard will be demonstrated and a suitable annotation tagset based on the TEI standard will be presented.

Sustainability of Richly Annotated Linguistic Corpora

Timm Lehmberg, M. A. (SFB 538 / University of Hamburg)

Dr. Georg Rehm (SFB 441 / University of Tübingen)

Dr. Andreas Witt (SFB 441 / University of Tübingen)

Max-Brauer-Allee 60, D-22765 Hamburg

tel: +49 40 42838-5309

fax: +49 40 42838-6116

email: tim.lehmberg@uni-hamburg.de

keywords: sustainability, annotated corpora, annotation standards

language: English

presenter: Timm Lehmberg

In this contribution, intermediate results from a sustainability initiative by three German research centres will be introduced. The project aims to create a sustainability architecture that provides long-term archiving as well as user-friendly querying of the large amount of linguistic corpora that has been collected and processed by the three centres.