Abs-109

Peter Menke (Universität Bielefeld, Germany) and Alexander Mehler (Goethe-Universität Frankfurt am Main, Germany)

From experiments to corpora: The Ariadne Corpus Management System

This paper describes the Ariadne Corpus Management System that assists researchers in all stages of generating multimodal speech corpora. It accompanies users from data acquisition via data analysis to the final publication of corpora. Ariadne is available to registered users in two variants:

1. as a web-based application that can be accessed from any computer connected to the Internet,
2. and as a client application that allows for a seamless integration with the user's file system structure.

Ariadne's field of application is characterized by tasks that scientists recurrently perform in research on multimodal communication:

- the encoding of observations elicited in experiments into a machine-readable format (transcriptions or annotations), possibly with the aid of different tools and their disparate data formats,
- the performance of various analyses on these data sets,
- the publication of subsets of resulting data as corpora in some interoperable format that is readable and further processable by others,
- and the compliance to privacy policies (e.g., anonymity of participants).

Ariadne assists at tasks from all these areas. First, it is built on top of a generic data model of communicative events, in combination with an expressive system of types. These types model various conditions and restrictions, and, when combined into special bundles, can accurately express the constraints of data formats from many popular third-party transcription and annotation tools. This mechanism helps to avoid unexpected alteration or fragmentation of data by specifically predicting what changes data will undergo when using a certain processing routine.

Data sets in the central format can then be processed by other modules inside Ariadne that perform part-of-speech tagging and lemmatization, calculation of measures of inter-annotator agreement, correction of values against value rules or vocabularies, and syntactic parsing. In addition, various functions from the field of statistics are provided -- either in the form of data preparation for input into third-party software or by performing analyses directly inside Ariadne.

One of the recent enhancements of Ariadne are components for the preparation and publication of corpora and related linguistic resources:

- Metadata modules for the collection of bundles of metadata which conform to different standards for the publication of language resources. These data sets make it easier for specialised search engines and crawlers to access and index resources properly.
- Interfaces and mechanisms for a flexible selection and publication of data to the public, in order to achieve compliance to the demands of various funding institutions.
- Closely related to these entrerprises is the goal of providing RDF functionality for complete corpora or resources and their fundamental components.

Our presentation and full paper will give a more exhaustive introduction to the functionality of the system. Examples of studies currently conducted in a research center on multimodal communication will serve as illustrations.

Gleim, R. & Mehler, A. (2010): Computational Linguistics for Mere Mortals --- Powerful but Easy-to-use Linguistic Processing for Scientists in the Humanities. Proceedings of LREC 2010, ELDA

Klyne, G. & Carroll, J.J. (2004): Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210

Menke, P. & Mehler, A. (2010): The Ariadne System: A flexible and extensible framework for the modeling and storage of experimental data in the humanities. Proceedings of LREC 2010, ELDA