LAUDATIO-Repository

Thomas Krause, Anke Lüdeling, Carolin Odebrecht, Laurent Romary, Peter Schirmbacher, Dennis Zielke

LAUDATIO-Repository

Accessing a heterogeneous field of linguistic corpora with the help of an open access repository

Thomas Krause, Anke Lüdeling, Carolin Odebrecht, Laurent Romary, Peter Schirmbacher, Dennis Zielke
Humboldt-Universität zu Berlin

An open access to digital historical research data for historical linguistics enables a fruitful exchange of research sources and research methods. To achieve this goal the LAUDATIO-Repository provides a long-term open access to historical corpus linguistic data. By developing the LAUDATIO-Repository we also want to explore how to build repositories that are useful for a set of well defined communities but are also flexible enough to be used and extended to serve other communities not considered beforehand. Considering the user community's needs requires a clear understanding of the community's user scenarios and research.

Digitalization and annotation of historical texts for linguistic purposes is a methodically and technologically challenging task that consumes a lot of time and resources. Linguistic analysis uses various annotations and methods such as multiple segmentations (Krause et al. 2012), normalizations (Bollmann, Petran & Dipper 2011), syntax trees (Nivre 2008), dependency annotations (Nivre et al. 2007) and (semi-) automatic consistency checks (Dickenson & Meurers 2003). Having the chance to either look up these methods and their results, the corpus itself, or to re-use existing corpora for further research may facilitate and enrich the research methods and possibilities of the linguistic community. Via an extensive documentation of the whole preparation phase including annotation, tools and the resulting data allows a uniform and structured access to such a heterogeneous field of research data and thereby a first establishment of best practice standards.

There are many tools and formats that are used to annotate and process historical corpora, e.g. for token based annotations and parses (see e.g. Schmidt 2009, Brugman & Russel 2004). That's why the repository is open to all formats. The repository is based on Fedora 3.6 (Lagoze et al. 2006) and a self-developed web interface that uses its API. Each format of a corpus is stored as whole in a separate binary data stream. The collection of data streams at a certain point in time form the corpus object. This mechanism is flexible enough to cover the variety of existing annotation tools such as EXMARaLDA, @annotate or ELAN and it can also cover new formats (even future ones not invented yet).

We also developed an unified meta-model for (historical) text corpora which is powerful enough to cover a heterogeneous field of corpus linguistic data and can be modified according to future developments in the research field[1]. For each corpus in the repository metadata in this meta-model must exist and on each import the metadata is automatically validated against the scheme. To enable a flexible modeling we chose a customization of TEI xml with the help of an ODD specification. The meta-model indirectly refers to basic concepts for using corpora not only

[1] The current documentation of the metadata can be accessed under http://korpling.german.hu-berlin.de/schemata/laudatio/doc/S6/corpus/

in the repository: A corpus is defined by the sum of documents, the actual (historical) texts. A document is the sum of all applied annotations. The mapping of this meta-model and the TEI header structure takes the user scenarios designed for the LAUDATIO-Repository into consideration: Every corpus need a structured and uniform documentation which comprises information about the corpus creation, the texts used for the corpus and every kind of annotation which is applied. Crucial for the reasons given above, the whole preparation process is covered as well.

In order to assist re-usage of existing corpus resources and minimize duplicate work researchers are able to find, download and import corpora on their own. The detailed metadata does not only support understanding the corpus, it is also used for searching for corpora in a structured manner and allows searching for corpora that are already suited for their research question or are a good base for further annotations. Thus, users are able to search for certain annotation methods, tools or annotation values. We use the metadata to provide a faceted search (see Tunkelang 2009 for an extensive discussion) using ElasticSearch[2] as a technical backend.

Via a flexible technical basis system and an extensive corpus documentation which is enable a structured view and search, an access to a heterogeneous field of historical corpora is provided by the LAUDATIO-Repository which will be presented in the demo session.

References

Bollmann, Marcel, Petran, Florian, Dipper, Stefanie (2011) Rule-Based Normalization of Historical Texts. *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop Hissar, Bulgaria, 16 September 2011.* pp. 34–42.

Brugman, Hennie, Russel, Albert (2004). Annotating Multimedia/ Multi-modal resources with ELAN. In: *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.*

Dickinson, Markus, Meurers, Detmar (2003) Detecting Errors in Part-of-Speech Annotation. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest, Hungary.

Krause, Thomas, Lüdeling, Anke, Odebrecht, Carolin, Zeldes, Amir (2012) Multiple Tokenization in a Diachronic Corpus. *Exploring Ancient Languages through Corpora Conference (EALC)*, 14.-16.Juni 2012.

Lagoze, Carl, et al. "Fedora: an architecture for complex objects and their relationships." *International Journal on Digital Libraries* 6.2 (2006): 124-138.

Nivre, Joakim (2008) Treebanks. *Corpus Linguistics: An International Handbook*. Vol. 1. Berlin: Mouton De Gruyter, 2008. pp. 225-41.

[2] ElasticSearch website, http://www.elasticsearch.org, accessed 29 Oct. 2013

Nivre, Joakim, Hall, Johan, Nilsson, Jens, Chanev, Atanas, Eryigit, Gülsen, Kübler, Sandra, Marinov, Svetoslav, Marsi, Erwin (2007) MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*. 13 / 02. pp. 95-135.

Schmidt, Thomas (2009) Creating and Working with Spoken Language Corpora in EXMARaLDA. In: Lyding, V. (ed.) *LULCL II: Lesser Used Languages & Computer Linguistics II*. pp. 151-164.

Tunkelang, Daniel (2009) Faceted search. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 1.1. pp. 1-80.