DGfS-CL

Christian Chiarcos*, Thomas Krause+, Anke Lüdeling+, Julia Ritz*, Viktor Rosenfeld+, Manfred Stede*, Amir Zeldes+ and Florian Zipser+

* Universität Potsdam
+ Humboldt-Universität zu Berlin

Search and Visualization of Richly Annotated Corpora with ANNIS2

This poster presents the latest version of ANNIS2, a web browser-based search and visualization environment designed to access richly annotated corpora with heterogeneous annotation schemes. Developed within Collaborative Research Centre 632 (SFB 632: "Information Structure: The Linguistic Means for Structuring Utterances, Sentences and Texts"), ANNIS (ANNotation of Information Structure) must meet the requirements imposed by diverse data from partner projects within the Research Centre and beyond.

Since information structure interacts with linguistic phenomena on many levels, the need to concurrently query and visualize data annotated for syntax, semantics, morphology, prosody, phonetics, referentiality and lexis, must be addressed, including where the data is multimodal. For this reason, ANNIS2 supports annotations of tokens, token spans and trees or other DAGs (directed acyclic graphs), and uses an appropriate query language capable of searching these structures. Both query language and visualizations are fully Unicode compatible to ensure support for a wide variety of non-European languages.

The underlying data for the system is annotated using both automatic taggers/parsers and a small set of manual annotation tools: EXMARaLDA (Schmidt 2004), annotate (Brants & Plaehn 2000) / Synpathy (www.lat-mpi.eu/tools/synpathy/), MMAX2 (Müller & Strube 2006), RSTTool (O'Donnell 2000) and PALinkA (Orasan 2003). These are then mapped onto the encoding standard of the SFB, PAULA (Potsdamer AUstauschformat für Linguistische Annotation / Potsdam Interchange Format for Linguistic Annotation), a stand-off multilevel XML format, which serves as the basis for further processing. The XML data is compiled into a relational database scheme, making the system's backend particularly scalable.

References

Brants T. & Plaehn, O. (2000) Interactive Corpus Annotation. In: *Proc. LREC 2000,* Athens.

Müller, C. & Strube, M. (2006), Multi-Level Annotation of Linguistic Data with MMAX2. In: Braun, S., Kohn, K. & Mukherjee, J. (eds.), *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt: Peter Lang, 197–214.

O'Donnell, M. (2000) RSTTool 2.4 – A Markup Tool for Rhetorical Structure Theory. In: *Proc. of the International Natural Language Generation Conference (INLG'2000), 13-16 June 2000*, Mitzpe Ramon, Israel, 253–256.

Orasan, C. (2003), Palinka: A Highly Customisable Tool for Discourse Annotation. In: *Proc. of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo.

Schmidt, T. (2004) Transcribing and Annotating Spoken Language with Exmaralda. In: *Proceedings of the LREC-workshop on XML Based Richly Annotated Corpora, Lisbon 2004*. Paris: ELRA.