

Comparing Integrated Linguistic Annotation Models

Maik Stührenberg

Universität Bielefeld

maik.stuehrenberg@uni-bielefeld.de

14. November 2007

Seamless integration of various, often heterogeneous (in terms of their output formats) linguistic resources and merging of the respective stand-off annotation layers are crucial tasks for linguistic research. After a decade of concentration on the development of formats to structure single annotations for specific linguistic issues, in the last years a variety of specifications to store multiple annotations over the same primary data has been developed. Especially the LINGUISTIC ANNOTATION FRAMEWORK (Ide and Romary (2004)) and the GRAPH-BASED FORMAT FOR LINGUISTIC ANNOTATIONS (Ide and Suderman (2007)) are of certain interest because of their origin as part of an ongoing international standardisation efforts (Ide and Romary (2007)). However, it is yet unclear, which impact these formats will have on the daily work of dealing with large linguistic corpora. In the project A2 “Sekimo” two specifications to store multiple annotated documents together with a collection of tools to analyse the data have been developed: the first format is based on a Prolog fact base and is discussed in detail in Witt (2004), Witt et al. (2005). The second format is XML based and uses a native XML database (a per-file-use is possible as well). Both formats can be used to analyse relations between elements of different annotation layers. While the Prolog fact base format relies on Prolog predicates, the XML based format uses standard XPath

and XQuery in contrast to the work described by Alink et al. (2006a,b), which involves new Standoff XPath axis steps. We will give an overview of both specifications developed in our project and will oppose them with the above mentioned formats for structuring linguistic corpora both in terms of flexibility and every day usage.

Literatur

- Alink, W., Bhoedjang, R., de Vries, A. P., and Boncz., P. A. (2006a). Efficient XQuery Support for Stand-Off Annotation. In *Proceedings of the 3rd International Workshop on XQuery Implementation, Experience and Perspectives, in cooperation with ACM SIGMOD*, Chicago, USA.
- Alink, W., Jijkoun, V., Ahn, D., and de Rijke, M. (2006b). Representing and Querying Multi-dimensional Markup for Question Answering. In *Proceedings of the 5th EACL Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*, Trento. EACL.
- Ide, N. and Romary, L. (2004). International Standard for a Linguistic Annotation Framework. *Journal of Natural Language Engineering*, 10(3-4):211–225.
- Ide, N. and Romary, L. (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hemsén, H., and Minker, W., editors, *Evaluation of Text and Speech Systems*, pages 263–284. Springer.
- Ide, N. and Suderman, K. (2007). GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the Linguistic Annotation Workshop*, pages 1–8, Prague, Czech Republic. Association for Computational Linguistics.
- Witt, A. (2004). Multiple hierarchies: New Aspects of an Old Solution. In *Proceedings of Extreme Markup Languages*.
- Witt, A., Goecke, D., Sasaki, F., and Lungen, H. (2005). Unification of XML Documents with Concurrent Markup. *Literary and Linguistic Computing*, 20(1):103–116.