DGfS-CL

Anna Hätty, Alisa Noha, Kerstin Eckart
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Metadata description: Addressing trainable components

Due to the amount and variety of linguistic resources, consistency and sustainability have become very important factors in linguistic research. It is therefore among the objectives of initiatives like CLARIN to ensure long-term availability and storage of linguistic data, cross-national networking and accessibility for users without deep technical background knowledge.

This abstract describes work done in the context of the CLARIN-D project, in particular at the CLARIN-D centre at the Institute for Natural Language Processing (Institut für Maschinelle Sprachverarbeitung, IMS), of the University of Stuttgart. We collect metadata of corpora and corpus tools, trainable tools and web services. The metadata is represented in the CMDI-format (Broeder et al. 2012), a flexible infrastructure which allows the user to create and reuse components for specific metadata aspects. These components can then be assembled into profiles, e.g. the *SpeechCorpusProfile* or the *TextCorpusProfile* created by the NaLiDa project which we use as a basis. The profiles support consistent documentation of different resources, e.g. the metadata instances of all text corpora described according to the *TextCorpusProfile* share the same basic structure which demands a minimum of information. If there is further information available, optional components can be used.

We focus on trainable tools such as the Tree Tagger (Schmid 1994). As these tools can be trained on different data sets, e.g. corpora differing in language, domain or annotation, there often exist several versions of one tool. For the metadata description it is therefore necessary to take the resources into account on which the tool has been trained. For our trainable tools, we separate the metadata description of the basic tool from the metadata description of its trainable components, such as parameter files or language models. For the trainable tool components we created the *ToolComponentProfile* by adapting the NaLiDa *ToolProfile*.

On our poster, we will discuss the questions “What are metadata?” and “Where do they come from?”. We will exemplify this by resources described at the CLARIN-D centre in Stuttgart, in particular by our profile for trainable tools. In addition, we will introduce other parts of the framework which helped us creating the metadata, e.g. ISOcat and the ComponentRegistry, and we will also show some examples for applications that make use of the metadata.

References

Broeder, D., Windhouwer, M., van Uytvanck, D., Goosen, T., & Trippel, T. (2012). CMDI: a Component Metadata Infrastructure. In *Proceedings of the workshop on Describing Language resources with Metadata. In conjunction with LREC 2012.*
Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing.*

http://www.clarin-d.de
http://www.clarin.eu
http://www.sfs.uni-tuebingen.de/nalida/
http://www.isocat.org/