Representation formats and models for lexicons

Thorsten Trippel
Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld

This presentation describes text-technological procedures of describing lexicons and offers a generic view on lexical information contained in lexicons. The lexicon here is the hypernym of dictionaries and lexicons from language and speech processing. Lexicons are generally described in terms of their microstructure (the structure of each individual lexicon article) and macrostructure (the order of the lexicon articles), sometimes also in terms of a mesostructure (the interrelation of lexicon articles and metadata). Text-technological approaches to modeling lexicons often start from the print dictionary, trying to describe the structure of the lexicographic information involved, using conventions of the Text Encoding Initiative (TEI), Expert Advisory Group of Language Engineering Standards (EAGLES), ISO standards for Terminology (ISO 12200, MARTIF). Others, more related to generative aspects of language describe their lexicons in terms of feature structures, modeled accordingly.

Feature structures are basically tree structures with some references, hence they can rather easily be modeled in XML using text-technological methodologies. Examples for these are the Draft International Standard ISO DIS 24610-1, which will allow the standardized representation of lexicons in this format. The TEI and EAGLES recommendations give explicit ways of structuring lexicon articles, originally in SGML but in the meantime in XML. Hence the structure of this lexicon formalism also relies on trees, and is already modeled in this paradigm. Computational applications and renderings of dictionaries on the web rely on the lexicon microstructure defining a table structure, using relational databases. This direct modeling of the different structures is full of redundancies and cannot easily be mapped on the XML structure. Redundancies can be avoided by normalization procedures according to the relational database theories laid out by Codd. Semantic ontologies finally are often rendered as trees, but taking a closer look even for ontologies this can hardly serve as a perfect formalism.

A generic model for the representing lexical information in a lexicon is presented with a graph structure, the Lexicon Graph Model. The graph model provides for the most generic way of modeling the relations of information units. A way of modeling and accessing the structure in the Text-technology paradigm is also provided.