Abs-220

Hanno Biber and Evelyn Breiteneder (Institute for Corpus Linguistics and Text Technology)

Corpus structures: The AAC Container as an example of organizing texts in a corpus

In the following paper corpora are regarded as complex containers of text. In order to be able to read, access and investigate the texts of a corpus in a digital environment, these containers must be constructed in a functional way. The corpus for which such a container is going to be developed is the AAC - Austrian Academy Corpus, a corpus of culturally and historically significant German language texts from the period between 1848 and 1989, which has been built at the Austrian Academy of Sciences in Vienna in the past years. More than 500 million running words of text have been scanned, converted into machine-readable text and annotated by means of XML-related standards. The AAC has collected thousands of literary objects and sources written by thousands of authors, representing an astonishing range of different text types. The texts that are systematically integrated into this large digital corpus have to be organized in a well structured way and provided with extensive metadata which is one important source of information about the texts. Specific metadata models and standards have been developed and become highly sophisticated methodologies for the description of corpora and their texts. The question of the structural organization of a large digital text corpus will be addressed from a new perspective of a large text corpus that has been built on the basis of specific thematic selection principles. In order to be able to investigate the texts and the language of the texts in such a large digital corpus of retrodigitized texts the corpus needs to be structured in a specific way so that the results given by means of applying analytical corpus linguistics methods and tools are useful for textual scholars in several ways. In making digital model editions of texts which are part of the overall corpus, some questions could be answered whereby the tools have become powerful and critical reading instruments equipped with fully searchable databases of the texts, with various indexes, search tools for lexicographic and linguistic research as well as navigation aids in a functional graphic design interface providing the reader with a complex research environment to read, study and access the texts from various entry points. Corpus research methods have to take into account the various textual representations of historical periods and their developments so that corpus based analysis of the texts are of value for linguistic, literary, and cultural studies as well as related fields.