Abs-222

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

500 000 000 tokens

In the following presentation an entire corpus of considerable size will be presented. The issue of representing a very large corpus in a format that offers only very limited space is paradigmatic for the general task of representing a language by just a small collection of texts and by just a small sample of the language. The AAC - Austrian Academy Corpus operated by the Institute for Corpus Linguistics and Text Technology consists of more than 500 million running words and several thousands of texts representing a wide range of different text types have been collected, digitized and annotated. Among the sources, which cover manifold domains and genres, there are literary journals, newspapers, novels, dramas, poems, advertisements, essays, travel accounts, cookbooks, pamphlets, political speeches as well as plenty of scientific, legal, and religious texts, to name just a few. The AAC was founded several years ago and is a corpus research initiative concerned with establishing and exploring large electronic text corpora and conducting scholarly research in the fields of digital text corpora. The texts that have been integrated into the collections of the AAC are German language texts of important historical and cultural significance. The historical period covered by the corpus is ranging from the 1848 revolution to the fall of the iron curtain in 1989. In this period significant historical changes with remarkable influences on the language and the language use in the German speaking areas can be observed and examined. The AAC corpus holdings provide a great number of reliable resources and interesting corpus based approaches for investigations into the linguistic and textual properties of these texts.