CLARIN-DK: Language data and language tools – the Danish node of an European Infrastructure

Lene Offersgaard, Mitchell John Seaton, Dorte Haltrup Hansen, Bart Jongejan
University of Copenhagen
leneo@hum.ku.dk, seaton@hum.ku.dk, dorteh@hum.ku.dk, bartj@hum.ku.dk

In the CLARIN-DK archive, the language based data you use for research purposes or want to share with fellow colleagues, can be stored. Currently, the archive contains Danish text corpora, both general language and language for special purpose, Danish audio, video and photo collections, lexica, wordnet, and linguistic annotations of texts and videos. The archive is one of the data archives in the European CLARIN infrastructure. Furthermore, CLARIN-DK includes a help desk and various tools. Danish corpora are available in the Korp concordance search tool, while the Voyant Tools allow distant reading of your own data or of pre-loaded Danish novels. In the CLARIN toolbox you can perform linguistic annotation of texts, create metadata and format your text into TEI. The linguistic tools included in CLARIN-DK comprise inter alia a tokenizer, a sentence segmentation tool, a part of speech tagger (PoS), a lemmatiser, an NP recognizer and a frequency analyzer.