Bulgarian National Corpus

Àâòîð Ñâåòëà Êîåâà

07.08.2014 ã.
Ïîñëåäíåå îáíîâëåíèå 08.08.2014 ã.

Ëåêöèÿ

We will discuss several key concepts related to the development of corpora and reconsider them in light of recent developments in Natural Language Processing. We propose a data-driven approach to corpus design, which integrates the best practices of traditional corpus linguistics with the potential of the latest technologies allowing fast collection, automatic metadata description and annotation of large amounts of data.

We will illustrate this concept with a description of the compilation, structuring, documentation, and annotation (morphosyntactic tagging, lemmatisation, word-sense annotation, annotation of noun phrases and named entities) of the Bulgarian National Corpus (http://ibl.bas.bg/en/BGNC_access_en.htm; http://ibl.bas.bg/en/BGNC_en.htm; http://search.dcl.bas.bg/). We will conclude with a brief evaluation of the quality of the corpus and an outline of its applications in Natural Language Processing and linguistic research.