Svetla Koeva, Diana Blagoeva, Sia Kolkovska, Tsvetana Dimitrova, Ivelina Stoyanova, Svetlozara Leseva
Institute for Bulgarian Language, Bulgarian Academy of Sciences
Sofia, Bulgaria

THE BULGARIAN NATIONAL CORPUS IN THE CONTEXT OF CONTEMPORARY LINGUISTICS

(Abstract)

The paper offers an overview of the methodology adopted in the development of the Bulgarian National Corpus (BuNC), its structure, linguistic annotation and applications with a focus on contemporary lexicography. BuNC is a dynamic corpus of over 5.4 billion words consisting of a monolingual (Bulgarian) part and 47 parallel corpora. The Bulgarian part includes about 1.2 billion words. The foreign language texts (originals or translations of texts in the Bulgarian part) total about 4.2 billion words, which makes BuNC the largest parallel corpus with a focus on Bulgarian. The corpus's structure is extendable and allows enrichment with new texts and categories in a way that ensures the corpus's representativeness and balance with respect to various features. The annotation approach adopted in the creation of the corpus involves automatic linguistic annotation on different levels: lemmatisation and PoS-tagging for Bulgarian and English, and sentence- and clause-alignment for parallel texts. BuNC can be accessed via a web application with advanced options for search using lemmas, wordforms and other linguistic (and extralinguistic) criteria, and functionalities for extraction of collocations and concordances. BuNC has been used for computational and theoretical linguistic research, in comparative studies and other areas. Special attention is paid to the corpus's applications in contemporary computational lexicography, with a focus on its use in the development of various dictionaries such as the comprehensive Dictionary of the Bulgarian Language, the Bulgarian FrameNet, the Bulgarian WordNet.

Keywords: Bulgarian National Corpus, parallel corpora, corpus design, corpus annotation, computational linguistics, computational lexicography

Published: 19 October 2015