Bulgarian Language Journal 58 (2011), 3, 5-6
ISSN 0005-4283

Svetla Koeva
Institute for Bulgarian Language, Bulgarian Academy of Sciences
Sofia, Bulgaria

CORPUS LINGUISTICS IN *БЪЛГАРСКИ ЕЗИК* (*BULGARIAN LANGUAGE*) JOURNAL

Corpus linguistics has significant applications and valuable achievements mainly in the field of lexicography and foreign language education. Computational linguistics also profits from large collections of row and annotated texts, for example extracting statistical evidences for linguistic phenomena. It has been proved long time ago that the Language technologies can effectively assist the corpus data processing to facilitate lexicographic work and grammar investigations. However, not all grammarians and lexicographers in Bulgaria assess the benefit from Natural language processing in increasing precision, completeness and effectiveness of scientific work (and even less of them – the advantages of effective collaboration with computational linguists).

The aims of the current issue of the Bulgarian Language journal are to present resent achievements in corpus analysis, to popularise the results of Bulgarian corpus linguistics and to attract attention to the large scale of applications offered by Natural language processing, corpus data analysis among them. The issue contains different types of articles – the leading article shows the efficient exploitation of computational linguistics techniques in the optimisation of lexicographic work (Rundell and Kilgarriff). The presented system – Sketch engine, processes also a Bulgarian annotated corpus (part of the Bulgariaн National Corpus) and uses Bulgarian word sketches. The structure, metadata description, Deep query language and some of the applications of Bulgarian National Corpus are described in the second article (Koeva, Blagoeva and Kolkovska). Next texts are devoted to Bulgarian speech corpus (Tisheva and Dzhonova) and Corpus of students' investigations (Aleksova, Laskova and Velkova) – both reporting significant results. The parallel corpora – corpora of translated documents in two or more languages which are sentence aligned – are used in different areas of application such as machine translation, multilingual information extraction, etc. Papers for two parallel corpora which contain Bulgarian counterparts: RuN-Euro corpus (Grønn and Rå Hauge) and Bulgarian-English-X+ language corpus (Koeva, Stoyanova and Dekova) are included in the issue as well. The parallel corpora studies are illustrated in several way: by comparative studies for Balkan languages (Tarpomanova), by more general (Blagoeva and Kolkovska) or specific investigations (Nestorova) based on the evidences taken from the Bulgarian National Corpus, as well as in the Language culture short papers (Blagoeva; Tarpomanova). The article focused on the diachronic corpora studies (Dimitrova) contributes to the variety of topics.

There are two reviews: for the *Bulgarian-Polish contrastive grammar* – the contrastive investigations are not feasible without empirical evidences from different languages (Laskowski)) and for *Bulgarian Sense-annotated Corpus* (Barkalova, Georgieva). The number ends with a chronicle for the *Third Corpus Linguistics Conference*, held in Spring 2011 (Kukova).

In general, the issue presents some of the basic achievements of the Bulgarian corpus linguistics in the compilation, annotation and exploitation of different type of linguistic corpora. At the same time the general tendencies and the future advances of Corpus linguistics are drown: compilation of very large corpora, reflecting a particular stage of the development of a given language in a given period; inclusion of different levels of detailed annotation representing appropriate linguistic information; wide usage of language technologies for corpora processing and results analysis in different research areas: lexicography, grammar, computational linguistics.

🖂 Svetla Koeva
svetla@dcl.bas.bg

Published: 30 September 2011