Bulgarian Language Journal 58 (2011), 3, 100-118
ISSN 0005-4283

Svetla Koeva, Ivelina Stoyanova, Rositsa Dekova
Institute for Bulgarian Language, Bulgarian Academy of Sciences
Sofia, Bulgaria

BULGARIAN-ENGLISH-X+ PARALLEL CORPUS

(Summary)

Following a short overview of the existing parallel corpora, which include Bulgarian, the paper focuses on the structure of the Bulgarian-English-X+ parallel corpus, the general principles of compiling and structuring parallel corpora, and the metadata used to describe the texts in them. Each subcorpus – Fiction, Law, News, Subtitles, Healthcare – is then presented in details: methods of compiling, languages included and the number of compiled words, the distribution of texts over thematic field, genre, and period of creation. Special attention is paid to the corpus annotation on various levels (morphological, morpho-syntactic, syntactic, and semantic). Due to the specificity of compiling this type of language resources (rather new for Bulgaria), the authors bring up also the question of the place of the parallel corpora with respect to the Copyright law.

Keywords: Bulgarian, English, parallel corpus, annotation, matadata

✉ Svetla Koeva
svetla@dcl.bas.bg
✉ Ivelina Stoyanova
iva@dcl.bas.bg
✉ Rositsa Dekova
rosdek@dcl.bas.bg

Published: 30 September 2011