Bulgarian Language Journal 58 (2011), 3, 119-130
ISSN 0005-4283

Tsvetana Dimitrova
Institute for Bulgarian Language, Bulgarian Academy of Sciences
Sofia, Bulgaria

DIACHRONIC CORPORA: PRELIMINARIES

(Summary)

The paper discusses the principles of compilation of diachronic corpora, while assessing the requirements for composition and structure of these collections of texts. The introduction outlines the tasks of diachronic corpus studies, along with a reveiw of the sources, namely the diachronic corpora on which observations were made. The first section gives an explanation on why historical linguistics is bound to rely on the approach of corpus linguistics. The second sections discusses the principles behind the diachronic corpora compiling, and a couple of issues in dealing with historical language data. The third section reviews the problems in structuring and processing of the language material prior to morphological annotation, during the stages of tokenization, lemmatization and tagset creation.

Keywords: diachronic language corpora, historical linguistics, tagset, linguistic annotation, corpus annotation, compilation, diachronic language variation

✉ Tsvetana Dimitrova
cvetana@dcl.bas.bg

Published: 30 September 2011