AGNĖ BIELINSKIENĖ, LOÏC BOIZOU, ERIKA RIMKUTĖ

LITHUANIAN MORPHOLOGICALLY ANNOTATED CORPUS AND TREEBANK

Summary

Annotated corpora are fundamental resources, which are very useful to develop language technology. The size, quality, and structure of such annotated corpora has a direct influence on the development of other tools. This article describes two annotated corpora prepared by the Centre of Computational Linguistics at Vytautas Magnus University: MATAS, a morphologically annotated corpus, and ALKSNIS, a treebank. It mainly discusses the structure and the tag set of both corpora, as well as the annotation procedure and tools. Both corpora are available online through ANNIS interface, therefore the syntax of ANNIS simple and complex requests is summarised for the Lithuanian potential users.

KEYWORDS: corpus, automatic morphological analysis, automatic syntactic analysis, treebank, language technologies.

AGNĖ BIELINSKIENĖ, LOÏC BOIZOU, ERIKA RIMKUTĖ
Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centras
V. Putvinskio g. 23–216, 44243 Kaunas
agne.bielinskiene@vdu.lt
erika.rimkute@vdu.lt
lboizou@gmail.com