

Automatic Text Mark-up Facilities Building Latvian Literature Corpus

NORMUNDS GRŪZĪTIS & KRISTĪNE LEVĀNE

IMCS, University of Latvia
29 Raina bulv., LV-1459, Riga, Latvia
normundsg@ailab.lv
kristine@ailab.mii.lu.lv

The Latvian Corpus at the Artificial Intelligence Laboratory of IMSC covers ca 30 mill. running words; ca 3.5 mill. running words are in the Latvian literature corpus, which is the part of corpus with free access on the web. This part is not copyright protected, and the corpus of the classics is interesting both for academic users and others. At the moment there are only simple navigation possibilities ensured on the web, so the main task of this project is to facilitate the use of literature corpus. The gained experience serves basis for the other software tools of Latvian Corpus, which are under the development.

From February, 2002 the development of Latvian literature corpus software tools has been carried out. The conception, requirements and the desirable tasks have been settled. So far, we have no common text structure standards and the content of corpus was HTML tagged. First, structure conception and standards based on XML technologies were created. Second, software tools and methods for the present corpus automatic transformation to the new build-up tagging system were developed. Presentation will deal with solutions and issues concerning this process.

DTD grammars are created for each Latvian literature genre (poetry, drama and prose). First DTD was made for poetry, because this genre is the most complicate. Different collections of poetry were examined, the aim was to try combining all the features in one grammar. In order to detect automatization problems, tagging tool was developed. For drama DTD, grammar by J.Bosak (<http://www.ibiblio.org/bosak>) for Shakespeare plays was used, which is a widely used example for drama structuring. The grammar for prose is relatively more simple. The current results of literature corpus transformation are available on www.ailab.lv/users/normundsg.

Next stage of the project is to create the whole corpus system and to develop software tools (navigation, concordance, statistics, and search) for end-users. Web interface will be provided giving the possibility to address wider audience and providing effective further development of the literature corpus.