The analysis of errors occuring during digitalisation of Hungarian text documents

Máté Pataki, Zoltán Tóth

Abstract

In the framework of the Meta-Contentum GVOP project MTA SZTAKI DSD examined what errors occur in scanned texts in Hungarian compared to their original ones, and searched for the roots of these errors. The results of the above analysis are elaborated in our presentation.

In our research we used a large quantity of real documents available digitally as well. The printed doc, rtf and txt documents were manually applied with errors, then scanned. The scanned versions were compared to the original texts so as to be able to check algorithmically the experiences gained earlier from the manual corrections.

The final database consisted one-gigabyte of text where we ran the comparative algorithms which, out of the 2 x 5500 test documents, compared the original and the corresponding scanned versions, collected the errors, statistics and words. The collection of words made further analysis possible, such as the percentage of unknown words in the database, the average rate of conjugated words with the same stem, rarely used words and word formations.

Link:

Departemnt of Distributed Systems, MTA SZTAKI: http//dsd.sztaki.hu