Îá îäíîì ìåòîäå àâòîìàòè÷åñêîé ãðàììàòè÷åñêîé ðàçìåòêè ñòàðîïå÷àòíûõ òåêñòîâ

Àâòîð Àðòåì Âèêòîðîâè÷ Àíäðååâ
07.09.2014 ã.

A
method is proposed for unsupervised morphosyntactic markup of old texts for
which no exact grammar nor vocabulary may be known. The method employs building
all possible mappings from text forms into grammemes and then reducing them
using a loose context-free (CF) grammar. The forms are further lemmatized based
on minimization of morphologic variation. The method has been tested on two old
Lithuanian documents from the late 16th century by M. Dauksha and has proven to
be rather efficient and accurate (up to 80 %).

andreev_elmanuscript2014 (594.8 kB)

Îá îäíîì ìåòîäå àâòîìàòè÷åñêîé ãðàììàòè÷åñêîé ðàçìåòêè ñòàðîïå÷àòíûõ òåêñòîâ