A corpus – based analysis of how accurately printed Romanian obeys to some universal laws

Adriana Vlad, Adrian Mitrea, and Mihai Mitrea
"POLITEHNICA" University of Bucharest
Faculty of Electronics and Telecommunications
1-3 Iuliu Maniu Bvd., Bucharest, Romania, vadriana@vala.elia.pub.ro

A main objective of the paper is how accurately printed Romanian complies with the stationarity hypothesis. A statistical approach to NL stationarity, based on the *m*gram structure is presented. The statistical inferences are: estimation theory with multiple confidence intervals, test of the hypothesis that probability belongs to an interval and test of the equality between two probabilities. The b size of the type II statistical error plays a special role in the designing of a corpus for mathematical purposes. The stationarity investigation was also used to investigate how accurately printed Romanian complies with two frequency–rank laws.

Key words: natural language stationarity, frequency–rank laws, multiple confidence intervals for probability.