Long Range Statistical Correlations in Human Language: A Case Study of Persian Language

Ali Mehri*

Department of Physics, Noshirvani University of Technology, Babol

Received: 09.04.2017 Final revised: 25.06.2017 Accepted: 30.12.2017

Abstract

The complex structure of human language enables us to exchange very complicated information. This communication system obeys some common nonlinear statistical regularities. We investigate four important statistical features of Persian language. We perform our calculations on six masterpieces from famous Persian scholars. Zipf's law and Heaps' law, which imply well-known power-law behaviors, are established in this language, showing a qualitative inverse relation with each other. Furthermore, the informational content associated with word ordering is measured by using an entropic metric. This metric can be applied in words relevance ranking process. We also calculate fractal dimension of words in the text by using box counting method. The fractal dimension of each word, that is a positive value less than or equal to one, exhibits its spatial distribution in the text. Generally, we can claim that the Persian language follows the mentioned statistical laws, like other languages studied in previous research.

Keywords: Text mining, Long range correlation, Zipf's law, Heaps' law, Entropy, Fractal dimension

* Corresponding author: alimehri@nit.ac.ir