Multifractal correlations in natural language written texts: Effects of language family and long word statistics

M. Chatzigeorgiou[1], V. Constantoudis[2], F. Diakonos[1], K. Karamanos[1], C. Papadimitriou[1], M. Kalimeri[3], H. Papageorgiou[4]

[1]Department of Physics, University of Athens, Greece

[2]Institute of Nanoscience and Nanotechnology, NCSR Demokritos

[3]Department of Physics, Tampere University, Tampere, Finland

[4]Institute of Language and Speech Processing, Athena R.C., Maroussi, Greece

During the last years, several methods from the statistical physics of complex systems have been applied to the study of natural language written texts. They have mostly been focused on the detection of long-range correlations, multifractal analysis and the statistics of the content word positions. During the last decades, several research groups have applied methods inspired by the statistical physics of complex systems to the detection and characterization of LRC and found that they are present in almost all mathematical representations of written texts at any scale of language hierarchy (letters, syllables, words, sentences, . . .) [1]. In the present paper, we show that these statistical aspects of language series are not independent but may exhibit strong interrelations [2]. This is done by means of a two-step investigation. First, we calculate the multifractal spectra using the word-length representation of huge parallel corpora from ten European languages and compare with the shuffled data to assess the contribution of long-range correlations to multifractality [3]. In the second step, the detected multifractal correlations are shown to be related to the scale-dependent clustering of the long, highly informative content words. Actually, we seek the link of the detected multifractal correlations to the positional scaling statistics of long words. We find that the clustering behavior of long words in real WLS presents an interesting crossover when we move to smaller inspection scales. In particular, at small scales a repulsing regime occurs where the long words repel each other and exhibit less clustering than the shuffled WLS. This crossover from clustering to anti-clustering behavior is responsible for the difference between real and shuffled multifractal spectra and dictates the footprint of LRC on multifractality. Furthermore, exploiting the language sensitivity of the used word-length representation and the benefits of our analyzed corpus (parallel texts in ten European languages), we demonstrate the consistent impact of the classification of languages into families on the multifractal correlations and long-word clustering patterns.

[l] M.A. Montemurro, D.H. Zanette, PLoS One **6**, e19875 (2011).

[2] M. Chatzigeorgiou et al., Physica A **469**, 173 (2017).

[3] P. Koehn, Proc. Conf. **The tenth Machine Tr**, 79 (2006).