Title: Complex systems perspective on natural language

Author(s): Stanisław Drożdż

Affiliation(s): (Faculty of Computer Science and Telecommunications, Cracow University of Technology and Institute of Nuclear Physics, Polish Academy of Sciences, Cracow, Poland

Keywords: Complex Systems, Natural Language, Correlations

Abstract : The science of complexity seeks to address the fundamental question of the governing principles that Nature employs when organizing the basic constituents of matter and energy into intricate structures and dynamic patterns that permeate all levels of the Universe's hierarchy. A closely connected phenomenon, namely natural language, has exhibited remarkable abilities in swiftly emerging and being adopted by humans. This linguistic phenomenon effectively mirrors these intricate patterns, as evidenced by its capacity to encode and communicate information pertaining to them and among them. Consequently, it is entirely justified to anticipate that natural language encapsulates the core essence of complexity. Indeed, this assertion holds particularly true in the context of human speech and writing, where the fact that *more is different* becomes strikingly evident. Therefore, it is only fitting for natural language to occupy a central role in the quantitative examination within the realm of complexity science.

Referring to such a perspective, this presentation [1] aims to consolidate the key methodological principles employed within this field and to assess their effectiveness in distinguishing between universal characteristics and language-specific traits within written representations of natural language across major Western languages. It thus delves into the examination of word frequencies in texts across major Western languages, highlighting the significant finding that accounting for punctuation largely restores the scaling behavior that is typically disrupted in the Zipf's law, especially for the most frequently used words, a phenomenon often addressed through the Mandelbrot correction.

Subsequently, the time series analysis techniques is utilised to investigate different forms of long-range correlations within written texts, drawing inspiration from complex systems. These time series are derived by segmenting the text into sentences or fragments between consecutive punctuation marks. Intriguingly, these series exhibit characteristics commonly observed in signals originating from complex systems, including the presence of long-range correlations and the emergence of fractal or even multifractal structures. Furthermore, a noteworthy observation is that the fluctuations in the distances between consecutive punctuation marks appear to universally follow the discrete Weibull distribution, a pattern often encountered in survival analysis.

Lastly, the utilization of complex network methodologies in the realm of linguistic structures, with a focus on word-adjacency networks is explored. These networks capture the relationships between words based on their co-occurrence within texts. The findings from such analyses suggest that the network metrics derived from these structures can serve as effective tools for tasks such as text classification, including authorship attribution. Complex networks have also been applied to a different category of linguistic networks known as word-association networks, which are constructed using data gathered from specific psycholinguistic experiments. Throughout all of the analyses presented in relation to written language, punctuation emerges as a pivotal factor, exerting a profound influence on the quantifiable characteristics of language.

[1] Based on collaboration with Jarosław Kwapień and Tomasz Stanisz