Łukasz Dębowski (Polish Academy of Sciences)

Natural Language Is Not A Finite-State Process:
Evidence from Three Statistical Power Laws

Abstract: Finite-state processes are hidden Markov processes with a finite number of hidden states. In the talk, we will argue that finite-state processes are insufficient as models of natural language. The idea is not new. It dates back to a famous debate between Burhus F. Skinner and Noam Chomsky (Skinner, 1957; Chomsky, 1959). Skinner believed that finite-state processes, with a sufficiently large number of hidden states can account for human language behavior. In contrast, Chomsky showed that if the syntax of natural language is described by a context-free grammar then natural language cannot be adequately described by finite-state formal languages or processes. Chomsky thought that rejecting finite-state processes means eradicating any probability models from linguistic considerations. In the talk, we will argue that this view is too radical. There exist stochastic processes which are not finite-state and we observe at least three different quantitative linguistic power laws which imply that natural language is not a finite-state process. The first power law concerns the growth of mutual information between two adjacent text blocks of increasing length (Hilberg, 1990), the second power law concerns the growth of maximal repetition in an increasing text fragment (Dębowski, 2012), whereas the third power law concerns the decay of mutual information between two letters separated by an increasing number of letters (Lin and Tegmark, 2017).