Finite-Size Scaling Law for Word-Frequency Distributions

Alvaro Corral[1,2,3,4] and Francesc Font-Clos[5]

[1] Centre de Recerca Matematica, Barcelona, Spain
[2] Departament de Matematiques, Universitat Autonoma de Barcelona, Spain
[3] Barcelona Graduate School of Mathematics, Barcelona, Spain
[4] Complexity Science Hub Vienna, Austria
[5] ISI Foundation, Torino, Italy

The dependence on text length of the statistical properties of word occurrences has long been considered a severe limitation for the usefulness of quantitative linguistics. We propose a simple scaling law for the distribution of word frequencies that brings to light the robustness of this distribution as text grows. In this way, the shape of the distribution is always the same, and it is only a scale parameter that increases (linearly) with text length. We give evidence for the validity of such scaling law, both using analytical arguments based on the generalized central-limit theorem and careful statistical tests.