

Ecological Models to Explain the Distributions of Words in Texts

Punchi-Manage R.¹✉, Mapa C.², Madhushani M.³, Dilhara M.¹, Karunathilaka P.⁴, Amiyangoda L.², Ekenayake U.²

Abstract

Genesis 1:1 says “*In the beginning was the Word, and the Word was with God, and the Word was God*”. It is found that less than 20% of the words can describe more than 80% of the contents of the word of a text. Pareto’s 80:20 rule, power laws, and zip’s laws are often used to explain the distribution of the words. However, we think that the ecological models can be also used to describe the distribution of the words. Species abundances of ecological communities are governed by a few dominant species followed by the majority of the rare and the singletons. This caused the species rank-abundance curves to show highly skewed distributions with long right tails; the patterns resemble the word distributions of texts. Ecologists often used three ecological models to explain species rank-abundance curves (*i.e.* Mac-Arthur’s Broken-Stick model, Fisher’s log-series model, and Preston’s Octave curves). The first step of our research is to use those three ecological models to see whether they could explain the word distributions of texts. For this purpose, we examined the relative frequencies of words in 10 renowned scientific literatures. We found that the relative frequency of word distributions of all the books was characterized each by a few dominant words preceded by a large number of rarely (infrequently) used words, hence causing long-tail distributions. We found Mac-Arthur’s Broken-Stick model and Fisher’s Log series model poorly explained the word distributions of texts. Also, the observed rank-abundances curves are outside the simulation envelopes of the Broken-Stick models. Further, Fisher’s log series models with different alpha values (parameters) could not explain the full pattern (high values explain only the tail distribution and low values explain only the dominant word frequencies). Interestingly, only Preston’s Octave curves are closely matched with observed relative word frequencies. Hence, our research emphasizes that the ecological model (*i.e.* Preston’s Octave curve) can be applied for statistical linguistics.

Keywords: *Log-series model, broken-stick model, octave curves, relative word distribution*

¹ Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya

² Department of Physical Sciences, Faculty of Applied Sciences, Rajarata University of Sri Lanka

³ Department of Mathematics, Faculty of Engineering, University of Moratuwa

⁴ Axiata Digital Labs Pvt. Ltd Level 11, Parkland Building, 33, Park Street, Colombo 02

✉ Corresponding Author: sarangaamila@gmail.com