

ON THE FRACTAL PATTERNS OF LANGUAGE STRUCTURES

Américo Tristão Bernardes*

Departamento de Física, ICEB/UFOP, Ouro Preto-MG, Brasil

Leonardo Costa Ribeiro[†]

*Departamento de Ciências Econômicas,
FACE/UFMG, Belo Horizonte-MG, Brasil*

Heliana Mello[‡]

Faculdade de Letras, UFMG, Belo Horizonte-MG, Brasil

(Dated: June 28, 2023)

Natural Language Processing (NLP) leverages AI algorithms to extract meaningful information from unstructured texts. In this study, we utilize NLP techniques to identify similar linguistic patterns across multiple languages. We employ the word2vec algorithm to generate 100-dimensional vector representations for English, Portuguese, German, Spanish, Russian, French, Chinese, Japanese, Korean, Italian, Arabic, Hebrew, Basque, Dutch, Swedish, Finnish, and Estonian. Subsequently, we calculate the fractal dimensions of the language structures, which exhibit multifractality with two distinct dimensions. These dimensions, combined with the token-dictionary size rate, enable us to represent the languages in a three-dimensional space. Analyzing the distances in this space reveals a correlation with the Phylogenetic tree, indicating the evolutionary relatedness between languages.

[1] Ribeiro LC, Bernardes AT, Mello H (2023) On the fractal patterns of language structures.

PLoS ONE 18(5): e0285630. <https://doi.org/10.1371/journal.pone.0285630>

* atb@ufop.edu.br

† lcr@cedeplar.ufmg.br

‡ heliana.mello@gmail.com