Menzerath and Menzerath-Altmann's law as a criterion of complexity in communication

Iván G Torre [1], Łukasz Debowski[2], and Antoni Hernández-Fernández[3]
[1]Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), San Sebastián, Spain
[2]Institute of Computer Science, Polish Academy of Sciences, Warszawa, Poland
[3]Complexity and Quantitative Linguistics Lab, Institut de Ciéncies de l'Educació, Universitat Politécnica de Catalunya, Barcelona, Spain,
[4]Societat Catalana de Tecnologia, Secció de Ciéncies i Tecnologia, Institut d'Estudis Catalans, Barcelona, Spain

Menzerath's law is a quantitative linguistic law which states that, on average, the longer is a linguistic construct, the shorter are its constituents. In contrast, Menzerath-Altmann's law (MAL) is a precise mathematical power-law-exponential formula which expresses the expected length of the linguistic construct conditioned on the number of its constituents:

$$y = \alpha m^{\beta} \exp(-\gamma m) \tag{1}$$

where α, β, γ are empirical parameters, being $\alpha > 0$, and usually $\beta < 0$ and $\gamma < 0$.

MAL and Menzerath's law have intensely been researched within the field of quantitative linguistics including written corpora, speech, and even music. Besides, similar analyses can be extended to any system that forms a hierarchy of units of different levels. Consequently, Menzerath's law has also been observed in genomes, protein domains, penguin vocalizations, chimpanzee gestural communication and primate vocalizations [1]. However, to fully consider whether MAL and Menzerath's law have an explanation from the point of view of compression and emerging complexity [2], we will do the exercise of studying them for so-called monkey typing models, which have previously heated the debate about other linguistic laws and brought stimulating results to the scientific community.

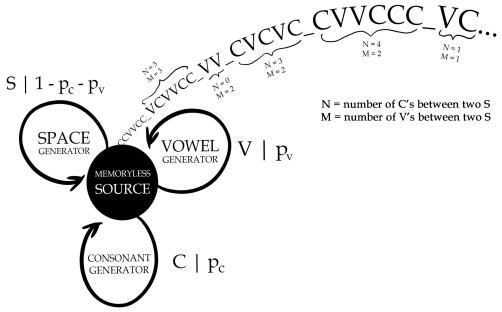

Fig. 1. The memoryless source model

Let N and M be the numbers of consonants and vowels, respectively, in a randomly generated word. Then, the mean length of a syllable equals $\frac{N+M}{M}$, if we assume that the number of syllables in the word can be approximated as the number of vowels M. Moreover p_c is the probability of emitting a C symbol in a one state memoryless source, p_v the probability of emitting V and $1-p_c-p_v$ the probability of emitting S (see Figure 1). Then, the exact form of Menzerath's law for the memoryless source is obtained:

$$E\Big(\frac{N+M}{M}\Big|M=m\Big) = \frac{a}{m} + b, \tag{2}$$

where

$$a = \frac{p_c}{1-p_c}, \quad b = 1 + \frac{p_c}{1-p_c}. \tag{3}$$

We show that this null model complies with Menzerath's law, revealing that Menzerath's law itself can hardly be a criterion of complexity in communication. This observation does not apply to the more precise Menzerath-Altmann's law, which predicts an inverted regime for sufficiently range constructs, i.e., the longer is a word, the longer are its syllables. To support this claim, we analyze MAL on data from 21 languages, consisting of texts from the Standardized Project Gutenberg. We show the presence of the inverted regime, not exhibited by the null model, and we demonstrate robustness of our results. We also report the complicated distribution of syllable sizes with respect to their position in the word, which might be related with the emerging MAL (see Figure 2). Altogether, our results indicate that Menzerath's law -in terms of correlations is a spurious observation, while complex patterns and efficiency dynamics should be rather attributed to specific forms of Menzerath-Altmann's law [3].

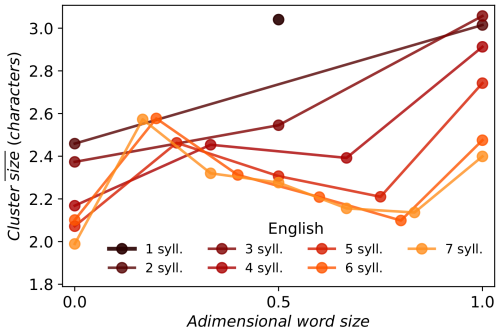

Fig. 2. Syllable sizes depending on word length and position

[1] S. Semple, R. Ferrer-i-Cancho and M. L. Gustison, *Linguistic laws in biology*, Trends in Ecology & Evolution (2021).

[2] I. G. Torre, B. Luque, L. Lacasa, C. T. Kello and A. Hernández-Fernández, *On the physical origin of linguistic laws and log-normality in speech*. R. Society open science, 6(8), 191023 (2019).

[3] I. G. Torre, Ł. Debowski and A. Hernández-Fernández, *Can Menzerath's law be a criterion of complexity in communication?*, Plos one 16(8), e0256133 (2021).