

The Distribution of Phonemic Contrasts across the World's Languages: Chance, Costs, and Integration across Linguistic Tiers

Fermín Moscoso del Prado Martín¹ (fm611@est.cam.ac.uk) & Suchir Salhan¹ (sas245@cam.ac.uk)

¹Department of Computer Science and Technology, University of Cambridge

In contrast with the broad corpus of research addressing the distribution of the frequencies of words across languages, the distribution of the frequencies of phonemic contrasts is remarkably understudied, with rare exceptions being Sigurd (1968), Martindale, Gusein-Zade, McKenzie, & Borodovsky (1996), Martindale & Tambovtsev (2007), and, most recently Macklin-Cordes & Round (2020). Using data from over one hundred language varieties, broadly distributed in typological, geographical, and genetic terms, we discuss that all of them can be modelled a priori by a virtually parameter-free distribution. Importantly, and in contrast with previous approaches, this distribution does not require any complex assumptions related to optimisation, preferential attachment, or the like. Rather, just two unquestionable facts of probability distributions (normalisation and symmetry) lead to a basically perfect model of the distribution of phonemic contrasts in any language. Using this distribution as a starting point, we demonstrate how the Principle of Maximum Entropy can be used to demonstrate how the plain marginal distribution of phoneme frequencies in a language contains important traces of its properties at other levels of description, including perceptual/articulatory factors, aspects of the specific language's phonotactics, and lexical information. Finally, we discuss how informational and diachronic constraints give rise to these distributions.

References

- Macklin-Cordes, J. L., & Round, E. R. (2020). Re-evaluating phoneme frequencies. *Frontiers in Psychology*, 11, 570895. <https://doi.org/10.3389/fpsyg.2020.570895>
- Martindale, C., Gusein-Zade, S. M., McKenzie, D., & Borodovsky, M. Yu. (1996). Comparison of equations describing the ranked frequency distributions of graphemes and phonemes. *Journal of Quantitative Linguistics*, 3(2), 106–112. <https://doi.org/10.1080/09296179608599620>
- Martindale, C., & Tambovtsev, Y. (2007). Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics*, 4(2), 1–11. http://www.skase.sk/Volumes/JTL09/pdf_doc/1.pdf
- Sigurd, B. (1968). Rank-frequency distributions for phonemes. *Phonetica*, 18(1), 1–15. <https://doi.org/10.1159/000258595>