Vowel-pair rank–frequency distributions are polylogarithmic

The distribution of phonemes in natural language has been shown to closely follow

$$(1) \quad f(r) = ar^{-b}c^{r}$$

where r is the phoneme's rank and $f(r)$ its relative frequency in the lexicon, and a, b and c are constants (Martindale et al. 1996, Tambovtsev & Martindale 2007). This formula, known as the polylogarithmic distribution (Kemp 1995),[1] belongs to a family of long-tailed distributions which arise in a number of non-linguistic areas including genetics, ecology, economics and sociology (e.g. Yule 1924, Champernowne 1953, Simon 1955, Chung & Cox 1994, Martindale & Konopka 1996); their occurrence across domains has been attributed to a general principle of "the rich get richer" whereby the probability of selecting an item from a set increases in proportion to its frequency in that set (Price 1976).

Within the context of language, which exhibits far more extensive combinatory properties than the domains mentioned above, the question naturally arises as to whether dependencies also follow a similar distribution. That is, if phonemes are described by (1), do combinations of phonemes exhibit the same, or at least a similar, rank–frequency distribution? And, if this is not the case, can the deviations from this base distribution be explained?

One relevant manner of combination in language concerns the selection of nuclei in adjacent syllables. Restricting attention to vowels for simplicity's sake, one may ask what the rank–frequency distribution of vowel pairs $x.y$ is, where vowel x occurs as the nucleus of the i^{th} and y as the nucleus of the $(i+1)^{\text{th}}$ syllable in a word. In order to investigate this, we constructed a genetically and areally diverse sample of 9 languages: Breton, Finnish, Georgian, Italian, Lozi, Malagasy, Northern Sami, Serbo-Croatian and Tagalog. For each language we consulted a digital or digitised dictionary consisting only of lemmata and no inflected forms; dictionary sizes ranged from 10,259 lemmata for Breton to 93,087 lemmata for Finnish. A custom-written R script was used to extract phonemic transcriptions from each dictionary and estimate the vowel-pair frequencies in the lexicon.

Our results show that vowel-pair frequencies do indeed conform closely to the polylogarithmic distribution (1): in each of the 9 languages studied, the empirical vowel-pair rank–frequency distribution is better fit by the polylogarithmic distribution than competing distributions, such as the Zipf distribution (Zipf 1949), the Sigurd distribution (Sigurd 1968) or the Borodovsky–Gusein-Zade distribution (Borodovsky & Gusein-Zade 1989), as illustrated in Fig. 1. When individual fits are examined in detail, however, it is possible to discern slight deviations from the theoretical distributions. These deviations can, in most cases, be traced back to phonological or morphological effects that skew the pair distribution away from what would be expected under purely random combination of elements. For example, Finnish exhibits vowel harmony that disallows—or at least strongly disprefers—combinations of front and back vowels (with the exception of neutral /i/ and /e/). This is reflected in the manner in which these pairs deviate from the theoretical polylogarithmic distribution (Fig. 1, inset).

The singleton rank–frequency distribution thus extends to pair rank–frequency distributions and describes at least some of the combinatorics inherent in language. This opens up the exciting prospect of modelling the interactions of phonotactics and language use in a way that derives the distribution in (1) as a mathematical prediction.

[1] For reasons unknown to us, Martindale et al. (1996), Martindale & Konopka (1996) and Tambovtsev & Martindale (2007) call this a "Yule distribution". In order to avoid a confusion with the Yule–Simon distribution, which is distinct from (1) but often also referred to as the "Yule distribution" (e.g. Simon 1955, Chung & Cox 1994), we choose to follow Kemp (1995) in calling (1) the polylogarithmic distribution.

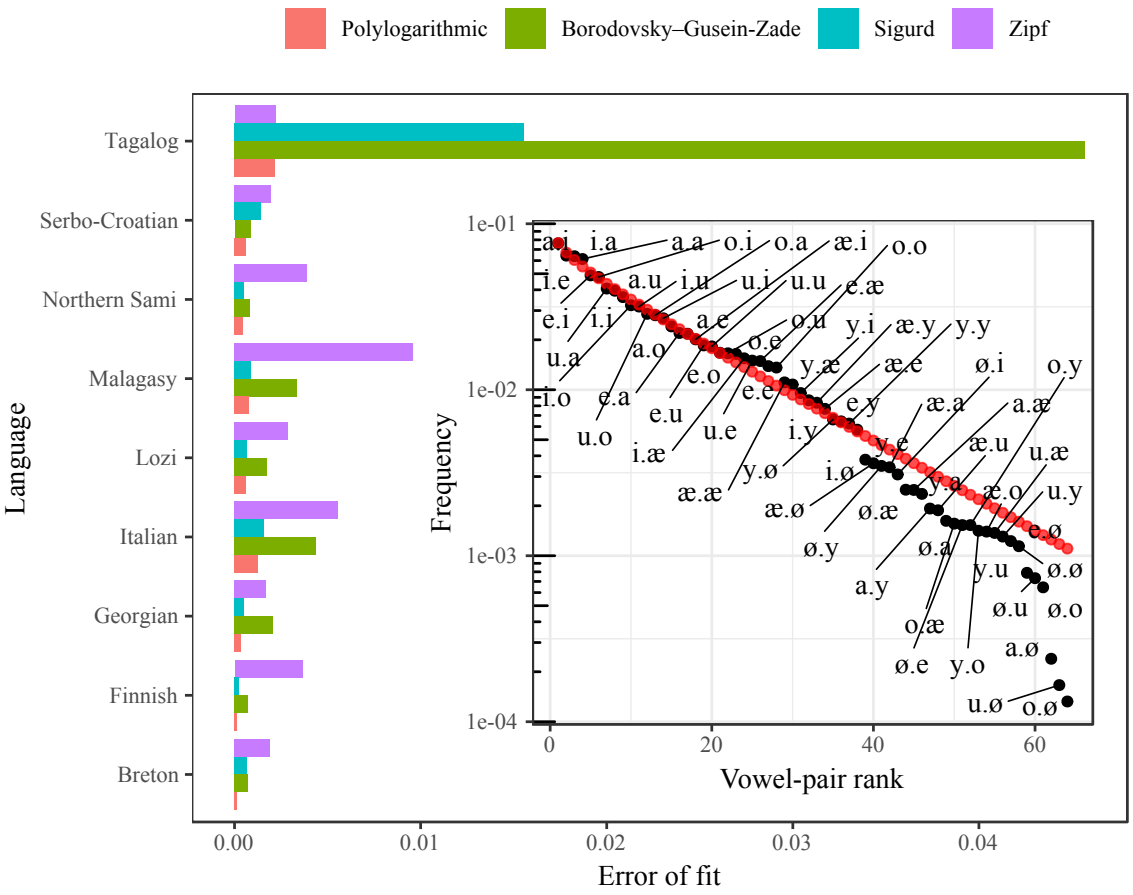

Figure 1: Empirical fit of various rank–frequency distributions to vowel-pair data in 9 languages. Inset: Fit of the polylogarithmic distribution (red) to Finnish data (black).

References

Borodovsky, Mark Yu. & Sabir M. Gusein-Zade. 1989. A General Rule for Ranged Series of Codon Frequencies in Different Genomes. *Journal of Biomolecular Structure and Dynamics,* 6. 1000–12.

Champernowne, David Gawen. 1953. A Model of Income Distribution. *The Economic Journal* 63(250). 318–51.

Chung, Kee H. & Raymond A. K. Cox. 1994. A Stochastic Model of Superstardom: An Application of the Yule Distribution. *The Review of Economics and Statistics* 76(4). 771–5.

Kemp, Adrienne W. 1995. Splitters, lumpers and species per genus. *Mathematical Scientist* 20. 107–18.

Martindale, Colin, Sabir M. Gusein-Zade, Dean McKenzie & Mark Yu. Borodovsky. 1996. Comparison of equations describing the ranked frequency distributions of graphemes and phonemes. *Journal of Quantitative Linguistics* 3(2). 106–12.

Martindale, Colin & Andrzej K. Konopka. 1996. Oligonucleotide frequencies in DNA follow a Yule distribution. *Computers & Chemistry* 20(1). 35–8.

Price, Derek J. de Solla. 1976. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* 27. 292–306.

Sigurd, Bengt. 1968. Rank-Frequency Distributions for Phonemes. *Phonetica* 18. 1–15.

Simon, Herbert A. 1955. On a class of skew distribution functions. *Biometrika* 42(3–4). 425–40.

Tambovtsev, Yuri A. & Colin Martindale. 2007. Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics* 4(2). 1–11.

Yule, George Udny. 1924. A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S. *Philosophical Transactions B* 213(21).

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley.