Physics Letters A ••• (••••) •••–•••

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Word frequencies: A comparison of Pareto type distributions

Martin Wiegand, Saralees Nadarajah, Yuancheng Si

School of Mathematics, University of Manchester, Manchester M13 9PL, UK

ARTICLE INFO

Article history:
Received 14 September 2017
Received in revised form 7 December 2017
Accepted 30 December 2017
Available online xxxx
Communicated by C.R. Doering

Keywords:
Kolmogorov–Smirnov test statistic
Squared error
Zipf's law

ABSTRACT

Mehri and Jamaati (2017) [18] used Zipf's law to model word frequencies in Holy Bible translations for one hundred live languages. We compare the fit of Zipf's law to a number of Pareto type distributions. The latter distributions are shown to provide the best fit, as judged by a number of comparative plots and error measures. The fit of Zipf's law appears generally poor.

1. Introduction

The primary means of communication among humans relies on the use of language to express ideas and emotions to one another. Depending on the language spoken, there appears to be a seemingly limitless amount of words. Strikingly certain words or word groups appear more often than others. This observation was first described by George Kingsley Zipf [23], who popularised an explanation based on the assumption that humans would use the most efficient way to describe a given concept.

Thus one would rather use specific, concise phrasing rather than a long-winded explanation with the same amount of informational value conveyed. Similar explanations had been mentioned earlier by Auerbach [4] and Jean-Baptiste Estoup as claimed by Manning and Schutze [15].

To quantify this assumption, Zipf provided a power law relationship between frequency and ranked usage. This relation can be applied to a number of naturally occurring sequence frequencies, such as medical or financial data [9], [5]. Mehri and Jamaati [18] applied Zipf's law to the word distribution of different languages based on one hundred translations of the Bible [13].

Zipf's Law and Pareto distributions are ubiquitous in language. They even exist when language is treated as networks: structural properties of weighted networks [17]; modeling in random texts [7]; structure-semantics interplay in complex networks [3]; statistical properties of unknown texts in the Voynich manuscript [2]; authorship recognition via fluctuation analysis of network topology [1].

We believe that the established method of a power law does not provide an appropriate fit and that the related Pareto type distributions could offer superior alternatives. After introducing the original formulation of Zipf's power law and applying it to the data, we do the same for the generalized Pareto, as well as Pareto type I-III distributions. We apply the Kolmogorov–Smirnov (KS) test statistic along with an R squared measure and a squared error. The different fits will be visualised by a number of comparative Log–Log plots for selected languages. Additionally we ran the fitting process on all languages stated in [18], and plotted the error measures accordingly, to visualise the effectiveness of both approaches. To verify the outcome for single author literary works we have added results on a number of different texts, as well as for randomly generated texts of different lengths (see [7]). We conclude this paper with a summary on different models and their suitability for further applications.

We would like to mention that there is a large body of work committed to understanding Zipf's law, more appropriate representations for rank frequency distributions, and why/when Zipf's law is broken. See [14], [10], [8], [12] and [21].

2. Pareto type distributions and Bible translations

From each translation of an identical Bible version a word frequency analysis is fashioned and words are ranked by use. Let

E-mail address: mbbsssn2@manchester.ac.uk (S. Nadarajah).

https://doi.org/10.1016/j.physleta.2017.12.056
0375-9601/

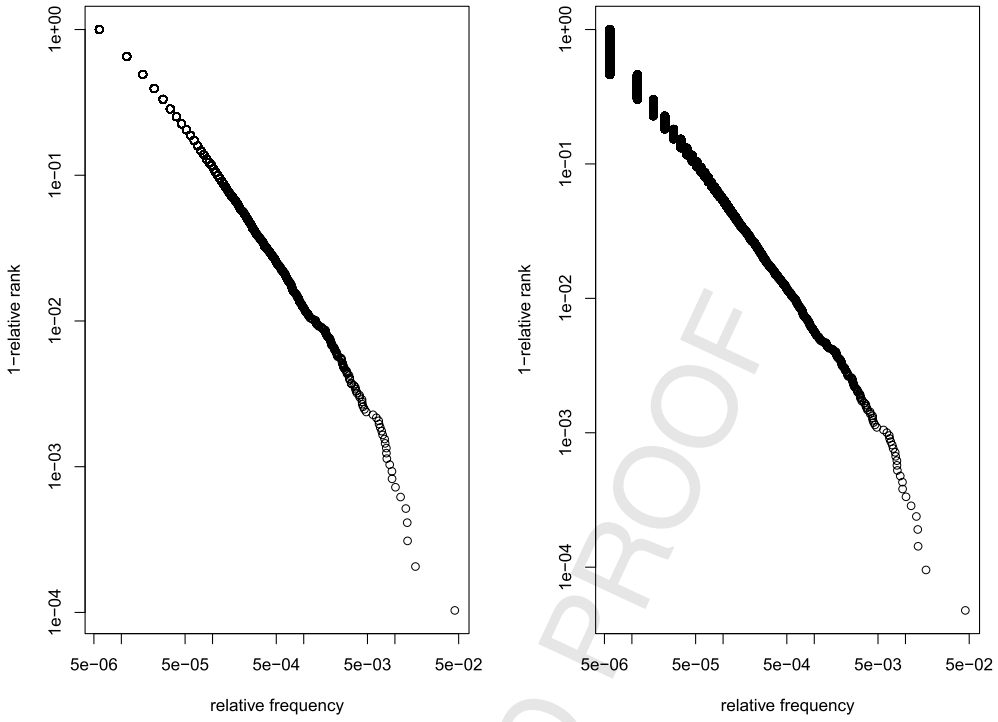

Fig. 1. A comparison for the achuar language using both ranking approaches.

N_v denote vocabulary size (number of used words) and N_t the text size (word count overall in the text). These are easily determined by tools such as [6], [22]; let r and f denote rank and frequency, respectively. The relative parameters are thus $r_r = r/N_v$ and $f_r = f/N_t$. At this point we like to note, that [18] was ranking the frequencies successively, meaning each rank was only given once and no two words could share the same value. This of course leads to a large amount of low-frequency words which occur only once or twice, covering a large range of ranks. This leads to the development of rank bands, as can be seen in Fig. 1. We believe this to be misleading, since given a large enough data set, words with the same frequency will display a difference in ranking in the hundreds or thousands. It is easy to see, how these bands will cause the deviation error to have a certain static base error, since no distribution will capture the entire band. This obscures the results, since the differences in goodness of fit would be miniscule. We have therefore chosen to allow multiple equal rank for equal frequencies, see the right hand side plot in Fig. 1.

We seek to establish a relation similar to Zipf's relation:

$$1 - r_r = \exp\left[a \log\left(f_r\right) + b\right],$$

where $a \in \mathbb{R}^-$ and $b \in \mathbb{R}$. Along with this formulation, we provide the performance results of the original relation provided by the Zipf law and the Zipf–Mandelbrot law:

$$\mathrm{freq}_{\mathrm{Zipf}}(r; s, N) = \frac{\frac{1}{r^s}}{\sum_{i=1}^{N} \frac{1}{i^s}},$$

$$\mathrm{freq}_{\mathrm{Zipf-M}}(r; s, q, N) = \frac{\frac{1}{(r+q)^s}}{\sum_{i=1}^{N} \frac{1}{(i+q)^s}},$$

where $r \in \mathbb{N}$ denotes the absolute rank and s, q denote the respective relation parameters.

As we will see in later plots, this relationship manifests itself as a straight line in a Log–Log plot. Especially in both lower and upper tails the distribution does not accurately capture the expected ranking. Below we have listed the cumulative distribution functions (CDFs) of the tested Pareto type distributions and the tested relationships between relative rank and frequency:

$$F_{P-I}(x) = 1 - \left[\frac{x}{\sigma}\right]^{-\alpha},$$

$$r_r = \left[\frac{f_r}{\sigma}\right]^{-\alpha},$$

$$F_{P-II}(x) = 1 - \left[1 + \frac{x-\mu}{\sigma}\right]^{-\alpha},$$

$$r_r = \left[1 + \frac{f_r-\mu}{\sigma}\right]^{-\alpha},$$

$$F_{P-III}(x) = 1 - \left[1 + \left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}\right]^{-1},$$

$$r_r = \left[1 + \left(\frac{f_r-\mu}{\sigma}\right)^{1/\gamma}\right]^{-1},$$

and

$$F_{PGPD}(x) = 1 - \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi},$$

$$r_r = \left[1 + \xi\left(\frac{f_r-\mu}{\sigma}\right)\right]^{-1/\xi}.$$

The parameters $\alpha > 0$, $\gamma > 0$ and $-\infty < \xi < \infty$ control the shape of these distributions. The parameter $\sigma > 0$ controls the scale of these distributions. The parameter $-\infty < \mu < \infty$ controls the location of these distributions. Smaller values of α correspond to