Word frequencies: A comparison of Pareto type distributions

Martin Wiegand, Saralees Nadarajah
School of Mathematics, University of Manchester, Manchester M13 9PL, UK

July 20, 2020

Abstract

Mehri and Jamaati [Physics Letters A, 381, 2470-2477, 2017] 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. A Pareto type III distribution is 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.

Keywords: Kolmogorov-Smirnov test statistic, Squared error, Zipf's law.

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 seems 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 bi-dimensional data, we do the same for the generalized Pareto, as well as Pareto types I-III distributions. We apply the Kolmogorov-Smirnov (KS) test statistic along with a 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 have 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 random ly generated texts of different lengths (see [7]). We will conclude this note 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 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 the original paper was ranking the frequencies succesively, 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 plot below). 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 (right hand side plot).

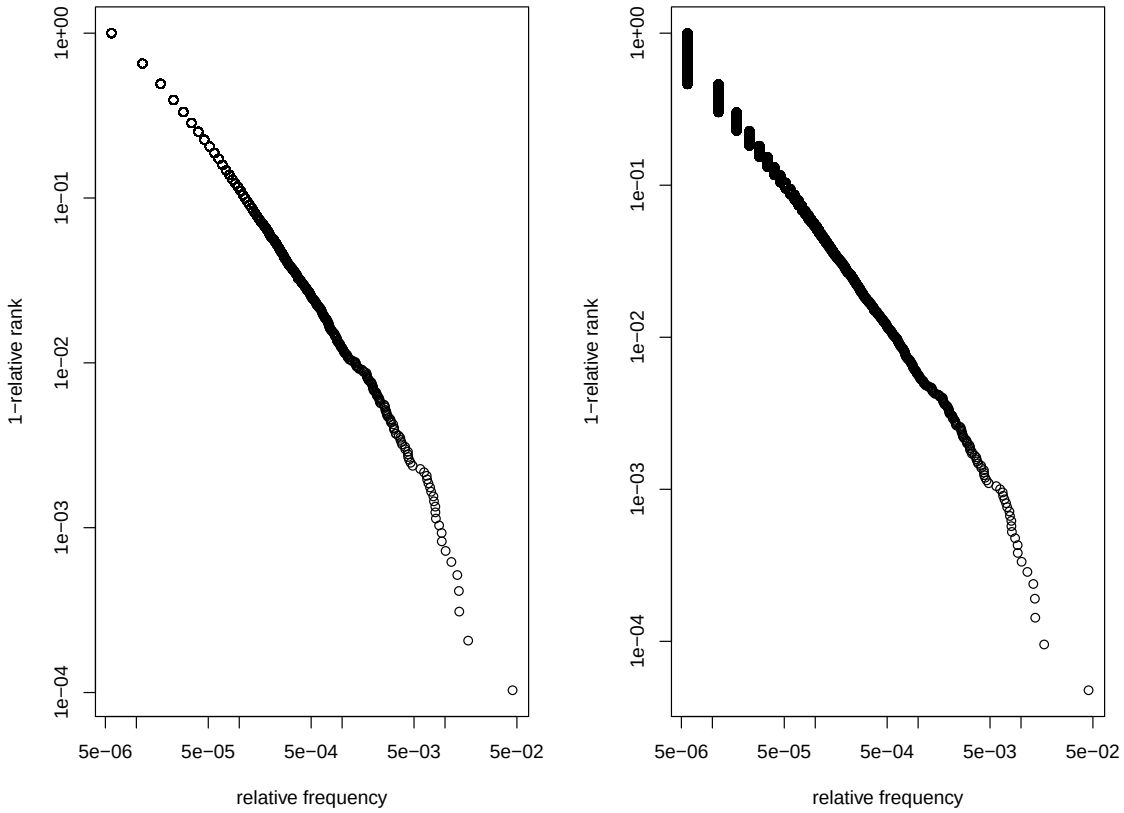

Figure 1: A comparison for the achuar language of both ranking approaches.

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 version given below.

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

Here $r \in \mathbb{N}$ is the absolute rank and s, q 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. Down below we have listed the cumulative density functions (CDFs) of the tested Pareto-type distributions and the tested relationships between relative rank and frequency:

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

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

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

and

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

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 heavier tails of the Pareto type I-III distributions. Larger values of γ correspond to heavier tails of the Pareto type III distribution. The generalized Pareto distribution has a finite tail if $\xi < 0$. It has an infinite tail if $\xi \geq 0$. The exponential distribution is the limiting case of the generalized Pareto distribution for $\xi \to 0$. The Pareto type II distribution is the particular case of the Pareto type III distribution for $\gamma = 1$. The Pareto type II distribution is a location scale variant of the Pareto type I distribution.

The four Pareto-type distributions and Zipf's law were fitted to the data by the minimization of the square deviation of the projected relative rank through a function of the relative frequency to the observed relative rank. Since the relation which was observed in the original paper was based on the CDFs of distributions, this was a more direct approach than density based approaches, say the MLE for example. We therefore provide the aggregated squared error, the Kolmogorov-Smirnov statistic as well as the R^2 measure on untransformed data, to diversify our measures. The squared error was minimized using the routine optim in the R software [19]. The routine uses a quasi Newton algorithm. The log-transformations are not considered at this point, that is during optimization. The logarithmic data is used solely to provide a better display of performances.

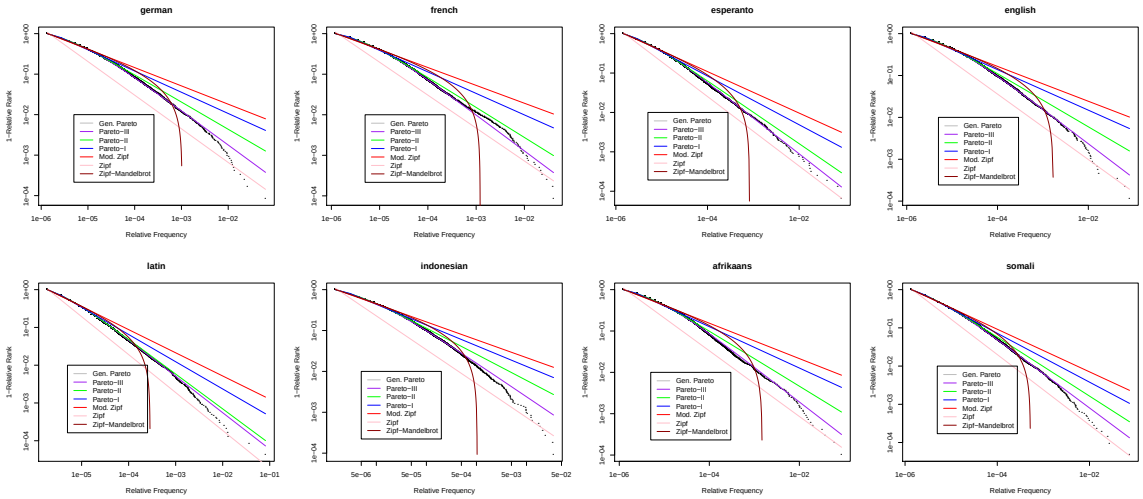

Figure 2: Log-Log inverse CDF plot of word relative frequency versus relative rank.

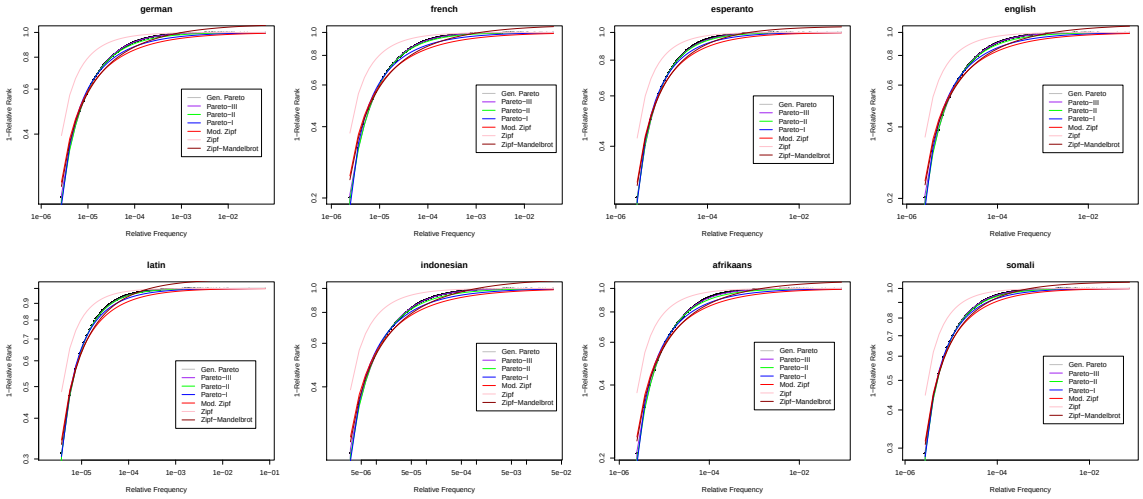

Figure 3: Log-Log CDF plot of word relative frequency versus relative rank.

The fit of the distributions for eight exemplary translations are shown in Figures 2 and 3. The picture is similar for all other languages. Zipf's method (red) does not adequately cover the curvature of the word distribution, whereas all other Pareto distributions offer a greater fit. The Pareto type III distribution repeatedly exhibits a squared error less than 0.5 thus providing the best overall performance, see Table 1, in Table 2 for further texts and in Table 3 for results by language family. Generalised and Pareto type II distributions have a slightly worse performance, yet seem to offer better results for certain languages (here Somali) and when considering the KS statistic. We have added the measure multiplied by the number of their parameters, to penalize for their complexity, yet the results remain the same, since the improvements of fit are to vast to be influenced by the different number of parameters.

	Distribution	KS statistic	Squared Error	R squared	Sq. Error x DoF
German	Gen. Pareto Dist.	0.012941	0.4847	0.9997585	1.4541
	Pareto Dist. Type III	0.007695	0.2879	0.9998566	0.8636
	Pareto Dist. Type II	0.012939	0.4847	0.9997585	1.4541
	Pareto Dist. Type I	0.036879	4.4943	0.9977609	8.9888
	Zipf-Mandelbrot Law	0.066079	7.8892	0.9960695	15.7785
	Zipf's Power Law (mod.)	0.062875	16.0795	0.9919891	32.1590
	Zipf's Power Law (orig.)	0.232340	415.5079	0.7929907	415.5079
French	Gen. Pareto Dist.	0.011481	0.2173	0.9998815	0.6517
	Pareto Dist. Type III	0.008465	0.2183	0.9998809	0.6518
	Pareto Dist. Type II	0.011485	0.2173	0.9998815	0.6517
	Pareto Dist. Type I	0.042907	5.8038	0.9968332	11.6075
	Zipf-Mandelbrot Law	0.058310	15.6081	0.9914835	31.21628
	Zipf's Power Law (mod.)	0.076270	25.2547	0.9862199	50.5094
	Zipf's Power Law (orig.)	0.236336	434.2414	0.7630579	434.2414
Esperanto	Gen. Pareto Dist.	0.010794	0.2820	0.9999024	0.8460
	Pareto Dist. Type III	0.005557	0.0264	0.9999909	0.0791
	Pareto Dist. Type II	0.010795	0.2820	0.9999024	0.8460
	Pareto Dist. Type I	0.032084	3.7774	0.9986926	7.5548
	Zipf-Mandelbrot Law	0.0457038	11.0022	0.9961921	22.0044
	Zipf's Power Law (mod.)	0.060283	18.0983	0.9937361	36.1966
	Zipf's Power Law (orig.)	0.195324	415.5079	0.8602944	403.6524
English	Gen. Pareto Dist.	0.015998	0.5933	0.9995604	1.7799
	Pareto Dist. Type III	0.008585	0.1792	0.9998672	0.5377
	Pareto Dist. Type II	0.015996	0.5933	0.9995604	1.7799
	Pareto Dist. Type I	0.043663	4.7441	0.9964850	9.4882
	Zipf-Mandelbrot Law	0.062942	7.9992	0.9940733	15.9984
	Zipf's Power Law (mod.)	0.070629	14.8486	0.9889985	29.6971
	Zipf's Power Law (orig.)	0.241482	326.2466	0.7582786	326.2466
Latin	Gen. Pareto Dist.	0.063925	0.1266	0.9999746	0.3797
	Pareto Dist. Type III	0.065341	0.0375	0.9999925	0.1125
	Pareto Dist. Type II	0.063925	0.1266	0.9999746	0.3797
	Pareto Dist. Type I	0.023692	3.1248	0.9993729	6.2496
	Zipf-Mandelbrot Law	0.067431	9.5414	0.9980852	19.0827
	Zipf's Power Law (mod.)	0.052449	20.6222	0.9958614	41.2445
	Zipf's Law (orig.)	0.187220	571.0397	0.8853997	571.0397
Indonesian	Gen. Pareto Dist.	0.016838	0.5856	0.9996690	1.7569
	Pareto Dist. Type III	0.010506	0.4492	0.9997461	1.3476
	Pareto Dist. Type II	0.016836	0.5856	0.9996690	1.7569
	Pareto Dist. Type I	0.039726	3.9837	0.9977487	7.9674
	Zipf-Mandelbrt Law	0.074594	6.0653	0.9965722	12.1306
	Zipf's Power Law (mod.)	0.065197	14.8299	0.9916190	29.6599
	Zipf's Power Law (orig.)	0.235931	412.3725	0.7669506	412.3725
Afrikaans	Gen. Pareto Dist.	0.019155	0.7884	0.9994930	2.3651
	Pareto Dist. Type III	0.009418	0.2494	0.9998396	0.7481
	Pareto Dist. Type II	0.019036	0.7882	0.9994931	2.3651
	Pareto Dist. Type I	0.047467	5.7356	0.9963114	11.4711
	Zipf-Mandelbrot Law	0.061464	9.8136	0.9936887	19.6273
	Zipf's Power Law (mod.)	0.074162	17.4008	0.9888094	34.8016
	Zipf's Law (orig.)	0.238492	356.1892	0.7709310	356.1892
Somali	Gen. Pareto Dist.	0.007878	0.1362	0.9999664	0.4086
	Pareto Dist. Type III	0.008510	0.4814	0.9998814	1.9255
	Pareto Dist. Type II	0.007877	0.1362	0.9999664	0.4086
	Pareto Dist. Type I	0.022617	2.2622	0.9994425	4.5244
	Zipf-Mandelbrot Law	0.047973	7.3160	0.9981971	14.6320
	Zipf's Power Law (mod.)	0.045634	14.6678	0.9963854	29.3356
	Zipf's Power Law (orig.)	0.186582	503.0458	0.8760347	503.0458

Table 1: Kolmogorov-Smirnov statistic, squared error value and the R squared value for the eight selected languages.

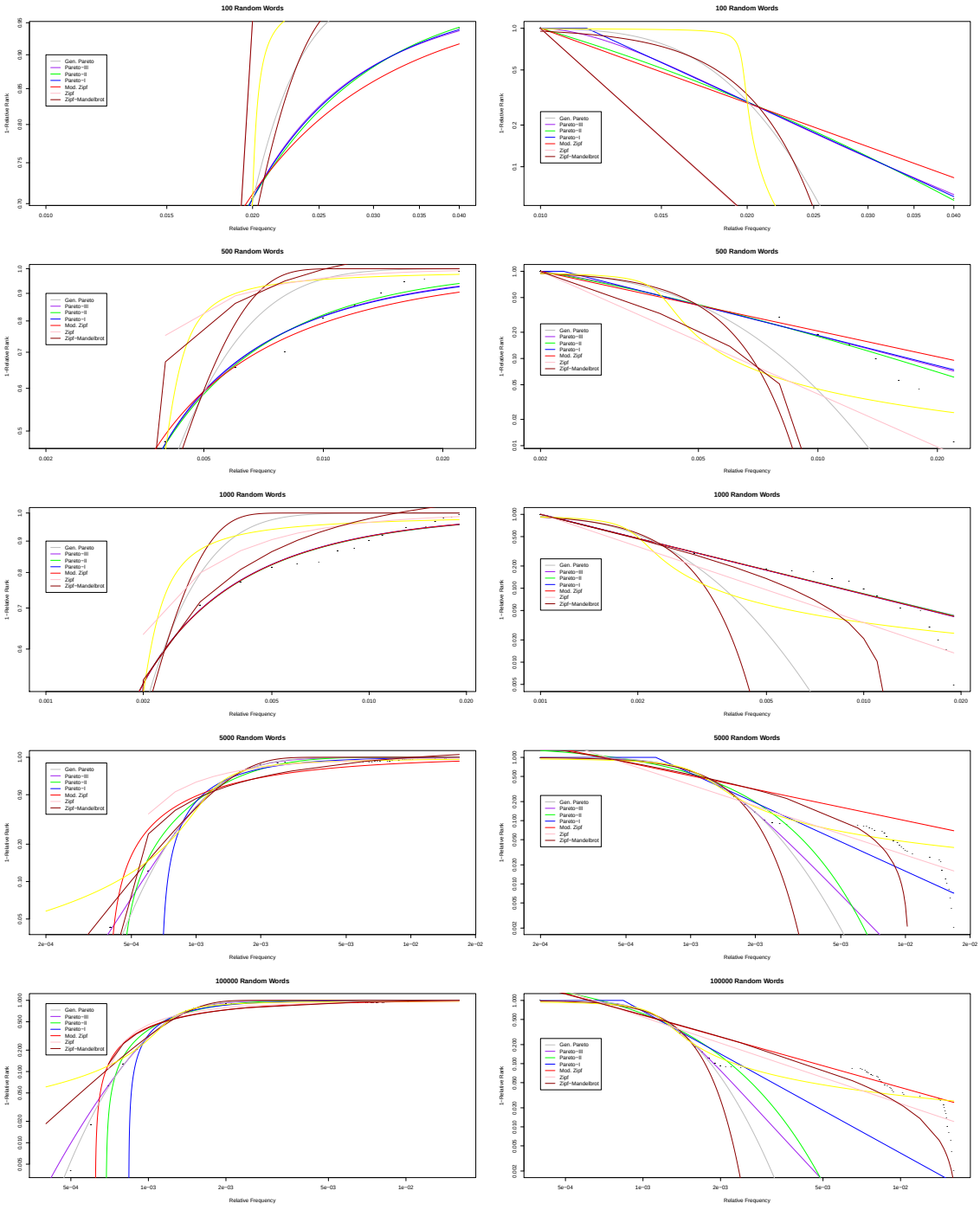

Figure 4: CDF and Inverse CDF for randomly generated texts of differents lengths.

The KS statistic, squared error value and R^2 value are shown in Table 1 for the eight selected languages. In all instances of the KS test of the Pareto type III distribution stays within the test limits. The standard Pareto distribution shows the largest deviation amongst the Pareto type distributions, yet still outperforms Zipf's law.

The plots in Figures 2 and 3 show the results for all one hundred languages, and provide further

Language	Distribution Characteristics		Das Kapital			The little Prince			Pinocchio		
	Distribution	DoF	KS Statistic	Squared Error	MSE times DoF	KS Statistic	Squared Error	MSE times DoF	KS Statistic	Squared Error	MSE times DoF
GERMAN	Pareto Dist. Type III	4	0.000989	0.101761	0.305	0.010283	0.014934	0.045	0.010920	0.054067	0.162
	Gen. Pareto Dist.	3	0.003131	0.044452	0.133	0.091233	0.023103	0.092	0.008274	0.013924	0.056
	Log Normal	2	0.125374	115.441400	230.883	0.125873	12.742430	50.970	0.125667	27.007040	108.028
	Pareto Dist. Type I	2	0.017812	0.691607	1.383	0.023751	0.094740	0.379	0.018940	0.104781	0.419
	Zipf-Mandelbrot Law	2	0.068001	2.018978	4.038	0.122776	0.802791	3.211	0.107907	0.898324	3.593
	Zipf's Power Law (Mod.)	2	0.040478	5.950488	11.901	0.042614	0.434373	1.737	0.035474	1.459127	5.837
	Burr	2	0.176045	211.143100	487.643	0.177032	23.013410	92.054	0.181333	49.486870	197.947
	Log Cauchy	2	0.135197	243.821600	274.126	0.281583	27.173910	108.696	0.123230	56.322290	225.289
	Zipf's Law (classic)	1	0.180526	274.125700	422.286	0.190018	28.772930	115.092	0.190238	73.065740	292.263
FRENCH	Pareto Dist. Type III	4	0.009758	0.338596	1.016	0.010826	0.024786	0.074	0.007784	0.026488	0.079
	Gen. Pareto Dist.	3	0.006765	0.083589	0.251	0.041232	0.028924	0.116	0.007600	0.027607	0.110
	Log Normal	2	0.112751	93.134930	186.270	0.138717	14.398440	57.594	0.105425	21.197170	84.789
	Pareto Dist. Type I	2	0.021963	1.098625	2.197	0.086805	0.060529	0.242	0.024887	0.299564	1.198
	Zipf-Mandelbrot Law	2	0.055796	4.431023	8.862	0.108780	0.762251	3.049	0.076005	1.017654	4.071
	Zipf's Power Law (Mod.)	2	0.047227	9.309323	18.619	0.035853	0.216583	0.866	0.053776	2.163238	8.653
	Burr	2	0.166082	173.945500	392.236	0.184818	25.873200	103.493	0.159304	40.450520	161.802
	Log Cauchy	2	0.111906	196.117800	299.774	0.296210	31.752670	127.011	0.269505	47.360050	189.440
	Zipf's Law (classic)	1	0.187909	299.773500	347.891	0.173797	23.529340	94.117	0.179127	55.309500	221.238
TURKISH	Pareto Dist. Type III	4	0.055628	0.003987	0.012	0.015492	0.055673	0.167	0.042636	0.026488	0.079
	Gen. Pareto Dist.	3	0.011121	0.068492	0.205	0.014462	0.047144	0.189	0.006587	0.018381	0.074
	Log Normal	2	0.127100	63.798500	127.597	0.158114	13.498670	53.995	0.107828	17.839130	71.357
	Pareto Dist. Type I	2	0.020189	0.164572	0.329	0.012170	0.017706	0.071	0.020472	0.130040	0.520
	Zipf-Mandelbrot Law	2	0.123286	1.140907	2.282	0.109375	0.737615	2.950	0.123267	0.976386	3.906
	Zipf's Power Law (Mod.)	2	0.034198	0.454144	0.908	0.029099	0.355008	1.420	0.049950	1.876398	7.506
	Burr	2	0.180358	89.186420	204.098	0.247408	24.305830	97.223	0.159509	34.397370	137.589
	Log Cauchy	2	0.129173	102.048800	126.688	0.169249	27.254480	109.018	0.110917	7.832272	31.329
	Zipf's Law (classic)	1	0.189347	126.687500	178.373	0.166607	20.661150	82.645	0.196562	61.574900	246.300
CHINESE	Pareto Dist. Type III	4	0.029207	0.092366	0.277	0.015158	0.030115	0.090	0.007770	0.018062	0.054
	Gen. Pareto Dist.	3	0.042020	0.260400	0.781	0.013652	0.028329	0.113	0.109392	0.059852	0.239
	Log Normal	2	0.070829	2.903704	5.807	0.163180	3.319680	13.279	0.265203	2.756889	11.028
	Pareto Dist. Type I	2	0.080017	1.466973	2.934	0.048063	0.637872	2.551	0.058623	1.648950	6.596
	Zipf-Mandelbrot Law	2	0.214256	0.524249	1.048	0.129672	1.047841	4.191	0.160274	1.782611	7.130
	Zipf's Power Law (Mod.)	2	0.108540	3.713127	7.426	0.086325	3.017451	12.070	0.095002	5.119466	20.478
	Burr	2	0.103798	6.982924	15.283	0.157066	7.734464	30.938	0.291168	7.256345	29.025
	Log Cauchy	2	0.120486	7.641737	67.054	0.170562	8.228197	32.913	0.291153	19.339500	77.358
	Zipf's Law (classic)	1	0.300499	67.053630	13.966	0.252295	44.228530	176.914	0.279496	69.237590	276.950
INDONESIAN	Pareto Dist. Type III	4	0.010335	0.026238	0.079	0.016569	0.023964	0.072	0.014447	0.023032	0.069
	Gen. Pareto Dist.	3	0.011806	0.025009	0.075	0.016388	0.027125	0.108	0.045285	0.017133	0.069
	Log Normal	2	0.124870	21.219420	42.439	0.310109	4.714009	18.856	0.105582	9.066435	36.266
	Pareto Dist. Type I	2	0.020668	0.117719	0.235	0.023566	0.060322	0.241	0.030346	0.166342	0.665
	Zipf-Mandelbrot Law	2	0.071161	0.246596	0.493	0.097035	0.265873	1.063	0.190379	1.030238	4.121
	Zipf's Power Law (Mod.)	2	0.035496	0.711013	1.422	0.037071	0.207894	0.832	0.060260	1.236936	4.948
	Burr	2	0.179344	37.597010	79.493	0.371639	9.852654	39.411	0.163023	17.619480	70.478
	Log Cauchy	2	0.313728	39.746680	45.496	0.421351	12.694100	50.776	0.538656	25.695250	102.781
	Zipf's Law (classic)	1	0.169070	45.495900	75.194	0.176325	7.731691	30.927	0.201514	34.707650	138.831
ENGLISH	Pareto Dist. Type III	4	0.008435	0.215744	0.647	0.073006	0.082995	0.249	0.009223	0.031210	0.094
	Gen. Pareto Dist.	3	0.010435	0.132124	0.396	0.072643	0.070940	0.284	0.008701	0.020349	0.081
	Log Normal	2	0.097387	63.291270	126.583	0.124206	9.123872	36.495	0.095324	11.657100	46.628
	Pareto Dist. Type I	2	0.028562	1.752500	3.505	0.014737	0.014279	0.057	0.031876	0.383924	1.536
	Zipf-Mandelbrot Law	2	0.066723	3.752940	7.506	0.085270	0.466470	1.866	0.124107	1.149552	4.598
	Zipf's Power Law (Mod.)	2	0.052403	8.702154	17.404	0.043061	0.878903	3.516	0.067266	3.063546	12.254
	Burr	2	0.150615	119.992500	262.432	0.181343	17.184570	68.738	0.151321	23.615840	94.463
	Log Cauchy	2	0.470600	131.215800	259.606	0.137855	18.049550	72.198	0.222663	79.252970	317.012
	Zipf's Law (classic)	1	0.204248	259.605600	239.985	0.170356	20.530430	82.122	0.216428	63.470480	253.882
RUSSIAN	Pareto Dist. Type III	4	0.003992	0.042233	0.127	0.079520	0.414370	1.243	0.044922	0.463791	1.391
	Gen. Pareto Dist.	3	0.005131	0.028837	0.087	0.083436	0.352813	1.411	0.051328	0.570696	2.283
	Log Normal	2	0.131363	213.313700	426.627	0.211210	6.684591	26.738	0.237524	39.735600	158.942
	Pareto Dist. Type I	2	0.012634	0.568300	1.137	0.065747	0.920781	3.683	0.057004	0.650418	2.602
	Zipf-Mandelbrot Law	2	0.072847	3.317209	6.634	0.324089	9.219785	36.879	0.140141	6.074273	24.297
	Zipf's Power Law (Mod.)	2	0.032475	8.037282	16.075	0.079103	2.263712	9.055	0.057002	0.650418	2.602
	Burr	2	0.179039	394.726000	925.695	0.150758	13.293000	53.172	0.315230	62.200950	248.804
	Log Cauchy	2	0.296138	462.847700	398.102	0.264715	20.097300	80.389	0.224976	79.252970	317.012
	Zipf's Law (classic)	1	0.162784	398.102200	789.452	0.337714	46.661530	186.646	0.215321	49.786120	199.144
SPANISH	Pareto Dist. Type III	4	0.003907	0.011462	0.034	0.009998	0.010424	0.031	0.095358	0.058764	0.176
	Gen. Pareto Dist.	3	0.009693	0.204077	0.612	0.007514	0.011410	0.046	0.093716	0.088388	0.354
	Log Normal	2	0.107839	96.216990	192.434	0.142627	13.028270	52.113	0.115551	25.078960	100.316
	Pareto Dist. Type I	2	0.022948	1.539986	3.080	0.014903	0.035024	0.140	0.023150	0.313913	1.256
	Zipf-Mandelbrot Law	2	0.041964	3.555577	7.111	0.087250	0.405491	1.622	0.066820	0.423785	1.695
	Zipf's Power Law (Mod.)	2	0.042811	7.157101	14.314	0.029674	0.207669	0.831	0.039450	1.223027	4.892
	Burr	2	0.159146	175.628500	390.770	0.197122	23.277600	93.110	0.175001	45.366850	181.467
	Log Cauchy	2	0.122077	195.384900	226.193	0.292608	27.608610	110.434	0.262715	51.498490	205.994
	Zipf's Law (classic)	1	0.175983	226.193100	351.257	0.162215	20.093140	80.373	0.174348	50.285390	201.142

Table 2: Error measures for three single author literary works.

evidence for what we had already assumed. . Please note that in this plot the generalized Pareto II and generalized Pareto distribution overlap, and show identical results in the comparative tables. This is due to the fact that both distributions can be emulated by each other by $\alpha = 1/\xi$ and

$\sigma = \sigma/\xi$, respectively. In Figure 5 notice that the Pareto type III distribution provides a tight grouping around 0 and 1 for the KS test statistic and the R squared measure, respectively. The Pareto type III distribution outperforms Zipf's law almost without fail.

As adressed by Ferrer-i-Cancho and Elvevå[7] we have observed the plausibility of the ability of Zipf's law to describe word frequencies. The results can be found in figure 4, as well as in a table in the Appendix. We can see an overall decline in the performance of all distributions, since the generated texts display distinct plateaus in their frequency distribution, which cannot be acurately covered by any distribution we have investigated.

However, we can confirm the findings of the referenced paper, and once again we can see similar behaviors of the distributions shaping up, as they did with the real life data. Especially the higher rates of simulation show how the Pareto-type distributions outperform the standard and modified Zipf laws.

To verify our findings, and prove the general validity of the conclusions we reach, we have repeated the procedure for a representative selection of different languages with additional single author texts, available in a number of translations (see [11],[20] and [16]).

The results are summarized in Table 2, listing the goodness of fit of every distribution investigated to each one of the literary works in question. We have tried to incorporate different subject matters and lengths of texts, to further diversify the type of text. Additionally the appendix offers similar comparative plots as for the bible translations.

For all three subsequent texts, we can see the same picture as for both the bible translations and random texts.

In Table 3 we have provided a data table of the results, grouped by language family along with a listing of the respective groupings. The overall results mirror our previous findings, but we can find aberrations for specific groups of languages. An example are the altaic languages, where the Pareto II distribution performs severly worse than the Zipf-Mandelbrot law. For for the japonic languages (japanese) and chibchuan we see a strong performance of the logarithmic distributions and the original Zipf law, which stand out distinctly from the remaining findings. These remain isolated cases though, and the Pareto III distribution specifically never completely fails to capture the behaviour of any given language translation.

	Pareto III		Pareto II		Pareto I		Log-Normal		Burr		Log-Cauchy		Zipf-Mandelbrot		Zipf (mod.)		Zipf (orig.)	
Family	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS
tai-kadai	0.065302	0.003905	0.016737	0.010281	0.781058	0.006824	11.487060	0.001752	543.921400	0.437138	11.600290	0.000007	151.336900	0.285007	0.781058	0.006826	151.342300	0.285010
sino-tibetan	5.068242	0.024913	3.927167	0.029489	4.548020	0.047344	187.446494	0.059947	724.560170	0.113205	333.873894	0.066933	178.183946	0.173669	15.873243	0.070492	643.672633	0.305140
japonic	0.018946	0.003123	0.001569	0.015856	0.066971	0.003066	0.165725	0.000009	73.886630	0.721251	0.165653	0.000005	3.535142	0.168982	0.066971	0.003064	3.536550	0.169014
basque	0.156630	0.008545	0.136694	0.009928	0.604888	0.023432	68.033820	0.075123	125.834400	0.107579	139.955400	0.098658	1.553431	0.096101	4.865370	0.044303	195.622400	0.185665
afro-asiatic	0.283393	0.009507	0.206428	0.010692	0.758778	0.015251	108.556720	0.086064	218.132269	0.121146	234.915004	0.098667	4.606928	0.097665	5.459691	0.033900	233.225489	0.171364
arawakan	0.009370	0.002996	0.016206	0.002483	0.046477	0.008969	78.797260	0.072113	187.280900	0.129435	172.757000	0.012430	1.141258	0.071912	1.449499	0.027142	105.036300	0.143394
iroquoian	0.405554	0.013269	0.402185	0.009065	0.414602	0.010478	107.784800	0.095982	229.759800	0.142486	214.326200	0.081926	0.989415	0.048020	0.455707	0.013352	89.347440	0.139589
dravidian	1.422745	0.334967	0.148587	0.017108	0.774644	0.352427	151.851610	0.388227	337.457747	0.407341	353.476007	0.412338	6.015302	0.078935	4.297268	0.361656	371.739650	0.197567
quechuan	0.025806	0.004399	0.039814	0.006743	0.565217	0.019794	62.668750	0.076391	116.306600	0.092716	127.386200	0.101962	1.477274	0.092096	4.439572	0.041852	174.344200	0.182790
uralic	1.342475	0.003346	0.136192	0.005012	1.739880	0.018118	197.371850	0.083800	395.141400	0.102263	426.396333	0.115003	5.421048	0.049767	11.473645	0.039835	394.455200	0.166132
algic	0.025976	0.003481	0.019568	0.003645	0.010523	0.003014	86.027440	0.087877	259.443345	0.178275	157.280580	0.009448	5.708840	0.103938	0.090348	0.007685	82.289570	0.149091
jivaroan	0.032366	0.005167	0.035617	0.005568	0.246817	0.012621	89.069743	0.085011	168.864267	0.116562	202.341467	0.092674	1.376517	0.058414	2.987229	0.030980	148.025993	0.144993
niger-congo	0.130750	0.003792	0.091129	0.008542	0.399769	0.016327	113.474365	0.083661	245.568573	0.113876	252.058948	0.105271	4.916187	0.072571	2.798109	0.032948	207.429338	0.174900
indo-european	1.101680	0.014092	0.984024	0.018624	3.322701	0.033752	117.547323	0.075931	214.246577	0.100760	243.684648	0.102248	6.861703	0.068882	11.808518	0.055665	330.276961	0.191888
constructed	0.026372	0.004063	0.282008	0.010371	3.777402	0.031718	100.953500	0.064536	195.809500	0.093275	215.251100	0.096269	11.002180	0.045704	18.098280	0.058735	403.652400	0.195324
equatorial	0.111452	0.005852	0.048002	0.006193	0.457512	0.016752	66.988660	0.075920	120.228200	0.108957	132.315600	0.095931	1.603569	0.029523	3.055777	0.032649	133.679800	0.159985
uto-aztecan	0.053420	0.003911	0.019017	0.003392	0.069841	0.007244	31.929350	0.083293	55.584840	0.101172	67.022850	0.121929	1.079570	0.025091	1.255562	0.025568	18.435360	0.086134
tucanoan	0.944945	0.000193	0.043409	0.003561	0.173757	0.011925	48.547140	0.072079	88.838370	0.104550	95.228510	0.094956	2.170866	0.031344	2.959732	0.034879	86.081120	0.140711
austronesian	1.852265	0.013659	1.257256	0.035925	2.228201	0.028665	76.591011	0.074512	137.786183	0.100597	155.182733	0.101208	6.634514	0.054628	10.170533	0.052553	238.815181	0.189547
altaic	0.739430	0.011218	49.646454	0.054695	3.324282	0.060736	92.794394	0.069873	194.010085	0.099586	237.832540	0.108679	13.412879	0.186525	27.993004	0.093638	578.716100	0.322971
austro-asiatic	4.561460	0.045128	11.855410	0.076074	16.119360	0.089358	26.977630	0.091200	47.903650	0.082134	55.234460	0.085804	7.814408	0.139316	16.386250	0.093848	321.037200	0.396108
oto-manguean	0.184014	0.013816	0.069150	0.014102	0.361999	0.024272	37.941917	0.080740	68.670555	0.102466	77.547960	0.103263	0.670217	0.041358	1.876963	0.038708	89.616470	0.188926
carib	0.217801	0.001357	0.190786	0.017052	0.843702	0.033886	24.475160	0.067208	45.526690	0.097894	47.923490	0.093342	1.173454	0.055765	2.917771	0.054351	99.661240	0.196119
west-papuan	0.096526	0.000007	0.017813	0.002130	0.142358	0.015625	15.563840	0.078274	40.934270	0.072608	29.992310	0.170894	1.347248	0.044067	1.823594	0.051011	32.730580	0.144357
nilo-saharan	0.303121	0.002407	0.223856	0.016738	1.497231	0.034246	31.019325	0.064340	59.249360	0.089631	62.064485	0.090875	2.286997	0.049582	5.219225	0.053364	174.189880	0.223889
chibchan	0.184356	0.032563	0.003739	0.004996	0.004843	0.005607	2.285363	0.084615	5.547749	0.125775	3.041124	0.098307	0.061381	0.016976	0.085436	0.018528	2.194719	0.090855
mayan	0.806186	0.043184	1.440334	0.043389	1.997668	0.055113	50.256480	0.083073	88.930968	0.103374	96.488878	0.106407	2.560421	0.074116	4.518950	0.072316	156.142997	0.195163
creole	0.290654	0.014703	0.261785	0.031554	0.663347	0.044242	8.886430	0.067522	16.696960	0.092612	15.248050	0.079270	0.235162	0.051215	1.310209	0.054700	61.631030	0.232228
DoF	3	3	3	3	2	2	3	3	2	2	2	2	2	2	2	2	1	1
Weighted Avrg.	**0.947075**	**0.023229**	**1.866162**	**0.019114**	**2.165788**	**0.038960**	**98.149189**	**0.085988**	**207.772701**	**0.125685**	**205.383934**	**0.105773**	**12.122039**	**0.077613**	**8.246670**	**0.058285**	**262.619623**	**0.191402**
AVG*DoF	2.841225	0.069686	5.598487	0.057343	4.331577	0.077920	294.447567	0.257964	415.545402	0.251369	410.767868	0.211547	24.244079	0.155226	16.493339	0.116570	262.619623	0.191402

Table 3: Distribution performances grouped by language family.

3 Conclusions

We have tested a number of distributions to accurately capture word frequencies. In comparison to the more established Zipf's law the Pareto type distributions have performed far better, supported by the Log-Log plots as well as the number of error measures we have calculated.

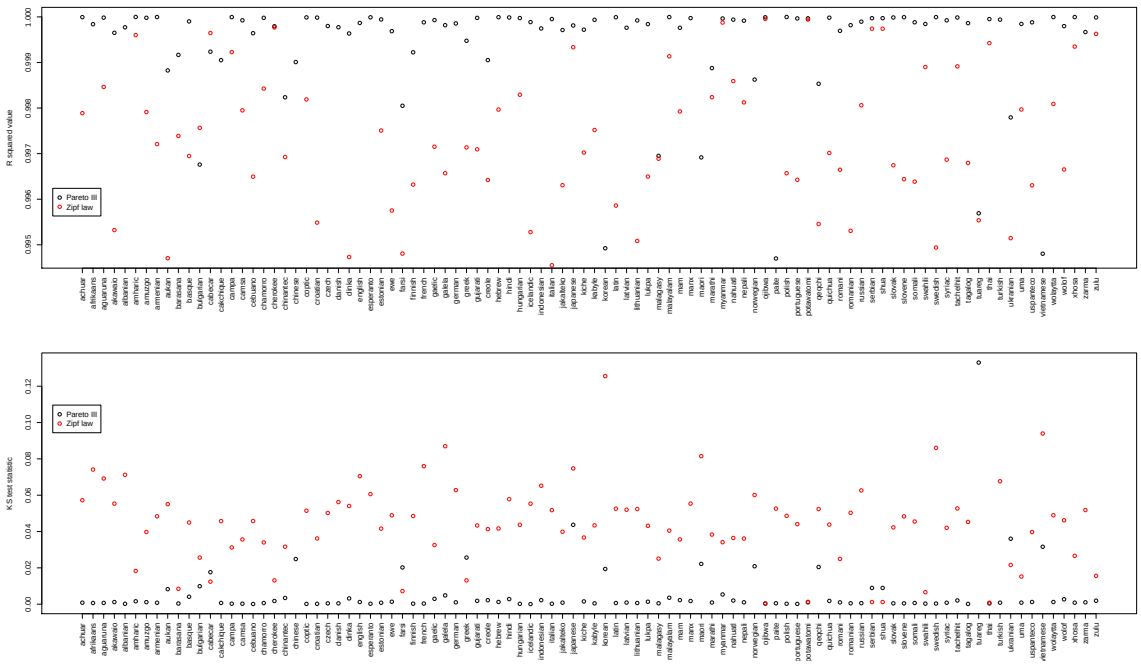

Figure 5: The K-S statistic and R squared measure for all languages compared.

The greater flexibility of the newly introdced distribution expectedly resulted in a better fit. Even though the Pareto distributions possess more parameters, the improvement of the fit is disproportionate to the increased complexity. The various incarnations of the Zipf law are in essence a rigid line on a Log-Log plot, which works somewhat for the body of the samples, but the Pareto-type distributions greatly improve the tail fit on boths ends of the frequencies.

It is evident to us that the models in this note are superior to Zipf's law and could greatly improve future modeling approaches.

Acknowledgments

We would like to express our thanks to Ali Mehri (University of Babol) for kindly providing the data sets used in [18] and for useful comments. The authors would also like to thank the Editor and the three referees for careful reading and comments which greatly improved the paper.

References

[1] D. R. Amancio, *Authorship recognition via fluctuation analysis of network topology and word intermittency.* Journal of Statistical Mechanics, 2015, doi:10.1088/1742-5468/2015/03/P03005

[2] D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira Jr, L. F. Costa, *Probing the statistical properties of unknown texts: Application to the Voynich manuscript.* PLoS ONE, 8, e67310, 2013, doi: 10.1371/journal.pone.0067310

[3] D. R. Amancio, O. N. Oliveira Jr, L. F. Costa, *Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts.* Physica A, 391, 4406-4419, 2012.

[4] F. Auerbach, *Das gesetz der bevolkerungskonzentration.* Petermanns Geographische Mitteilungen, 59, 74-76, 1913.

[5] R. H. Baayen, *Word Frequency Distributions.* Springer Science and Business Media, 2002.

[6] *https://www.browserling.com/tools/word-frequency*

[7] R. F. Cancho, B. Elvevag, *Random texts do not exhibit the real Zipf's law-like rank distribution.* PLoS ONE, 5, e9411, 2010, doi: 10.1371/journal.pone.0009411

[8] R. F. Cancho, R. V. Sole, *Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited.* Journal of Quantitative Linguistics, 8, 165, 2001.

[9] R. F. Cancho, R. V. Sole, *Least effort and the origins of scaling in human language.* Proceedings of the National Academy of Sciences of the United States of America, 100, 788-791, 2003.

[10] A. Clauset, C. R. Shalizi, M. E. J. Newman, *Power-law distributions in empirical data.* SIAM Review, 51, 661, 2009.

[11] C. Collodi, *Pinocchio* First Edition, 1883.

[12] M. Gerlach, E. G. Altmann, *Stochastic model for the vocabulary growth in natural languages.* Physical Review X, 3, 021006, 2013.

[13] *http://homepages.inf.ed.ac.uk/s0787820/bible/*

[14] B. B. Mandelbrot, *An informational theory of the statistical structure of languages.* In: Communication Theory, pp. 486-502, 1955.

[15] C. D. Manning, H. Schutze, *Foundations of Statistical Language Processing.* MIT Press, 1999.

[16] K. Marx *Das Kapital. Kritik der politischen Ökonomie* Volume I, 1867.

[17] A. P. Masucci, G. J. Rodgers, *Differences between normal and shuffled texts: Structural properties of weighted networks.* Advances in Complex Systems, 12, 2009, doi: 10.1142/S0219525909002039

[18] A. Mehri, M. Jamaati, *Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations.* Physics Letters A, 381, 2470-2477, 2017.

[19] R Development Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017.

[20] A. de Saomt-Exupery, *Le Petit Prince.* 1947.

[21] J. R. Williams, J. P. Bagrow, C. M. Danforth, P. S. Dodds, *Text mixing shapes the anatomy of rank-frequency distributions.* Physical Review E, 91, 052811, 2015.

[22] *http://www.writewords.org.uk/wordcount.asp/*

[23] G. K. Zipf, *Human Behavior and the Principle of Least Effort.* Addison-Wesley, Massachusetts, 1949.

Appendix

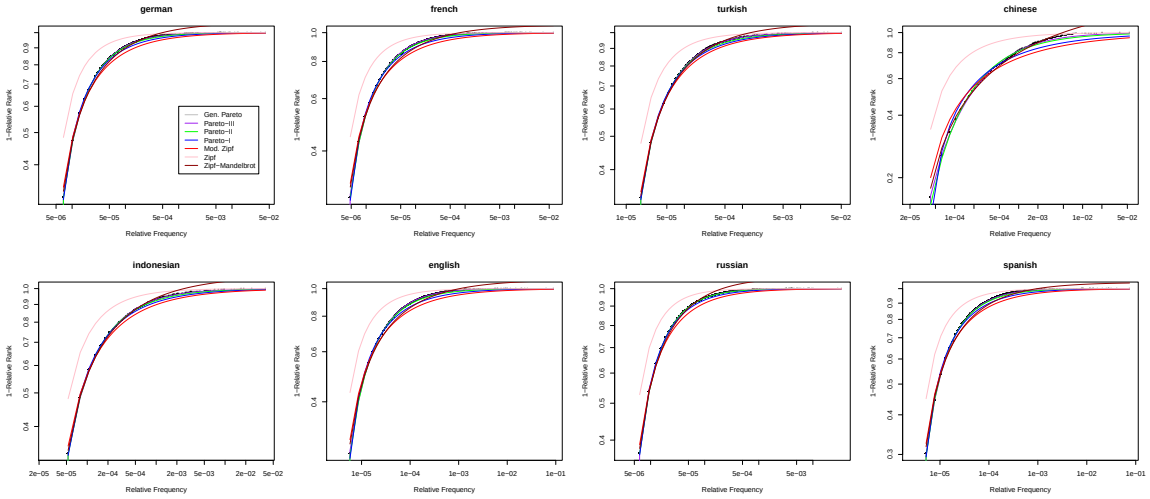

Figure 6: Log-Log CDF plot of word relative frequency versus relative rank for "Das Kapital".

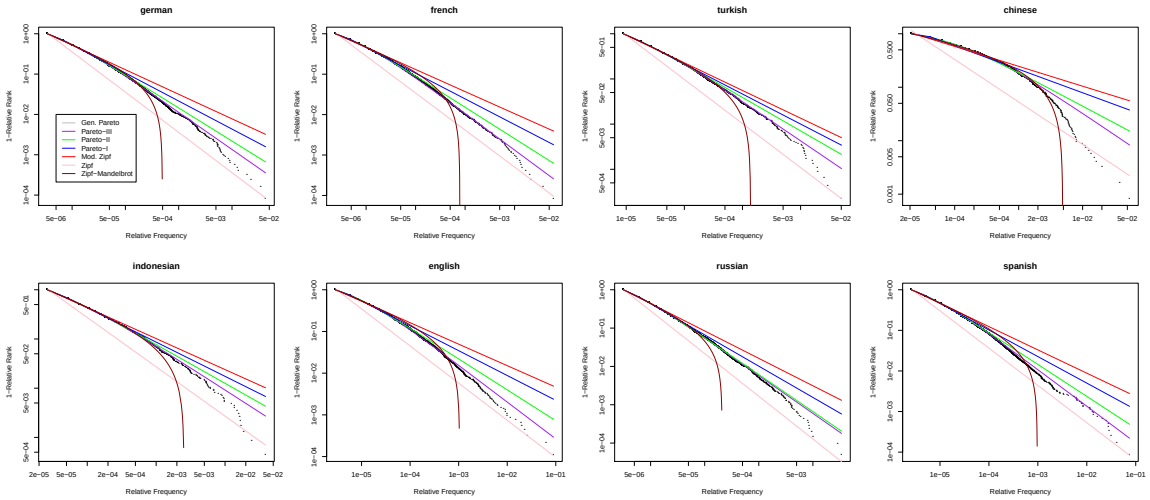

Figure 7: Log-Log CDF plot of word relative frequency versus inverse relative rank for "Das Kapital".

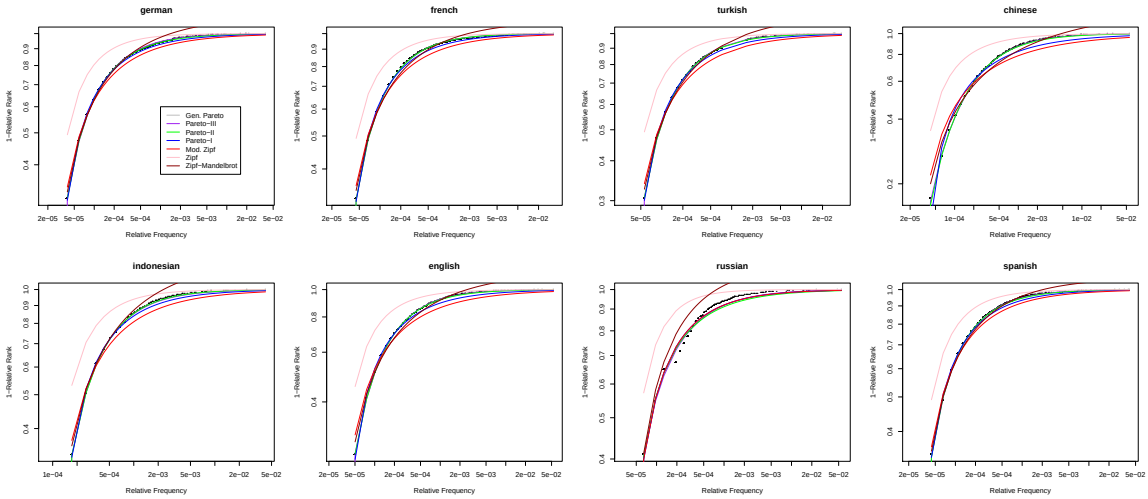

Figure 8: Log-Log CDF plot of word relative frequency versus relative rank for "Pinocchio".

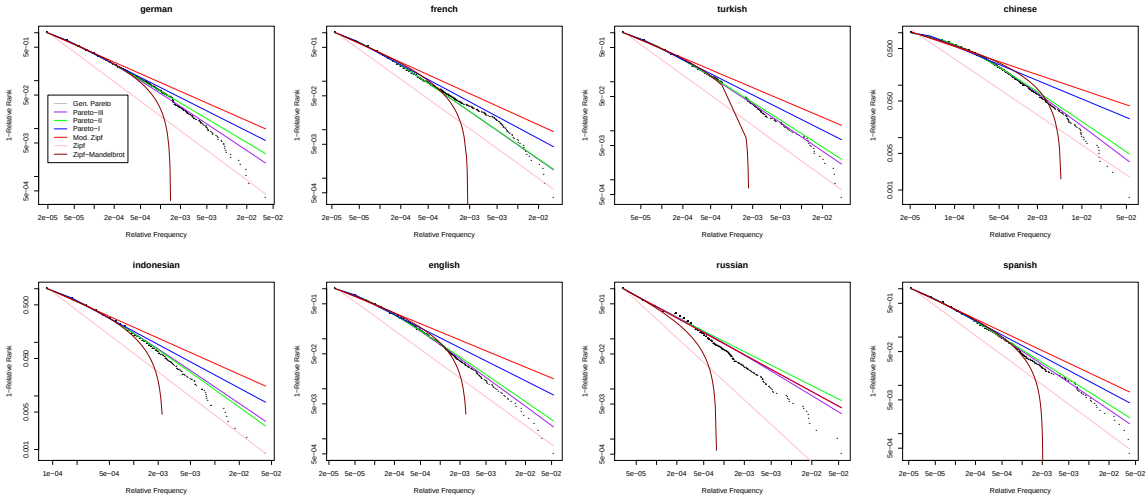

Figure 9: Log-Log CDF plot of word relative frequency versus inverse relative rank for "Pinocchio".

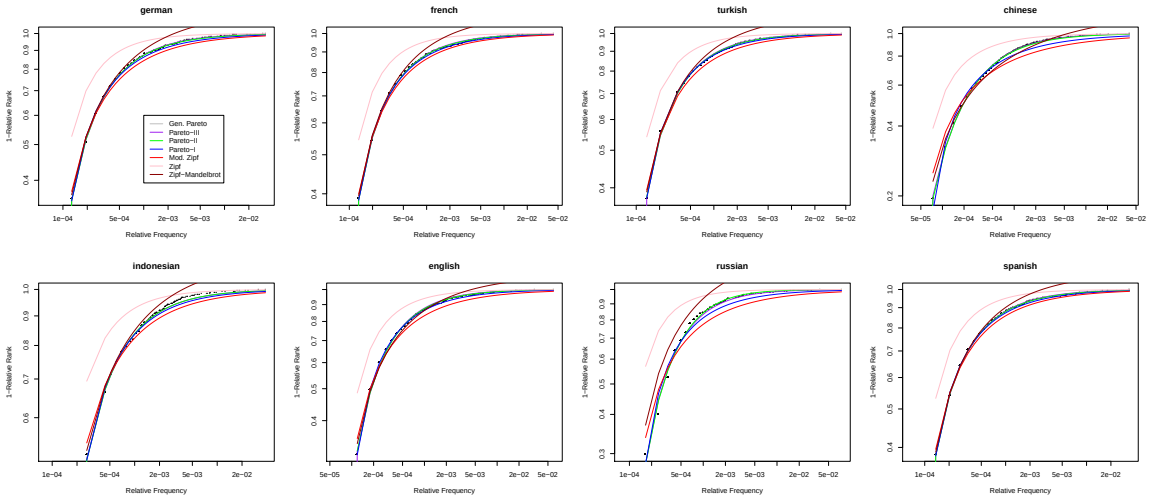

Figure 10: Log-Log CDF plot of word relative frequency versus relative rank for "The little Prince".

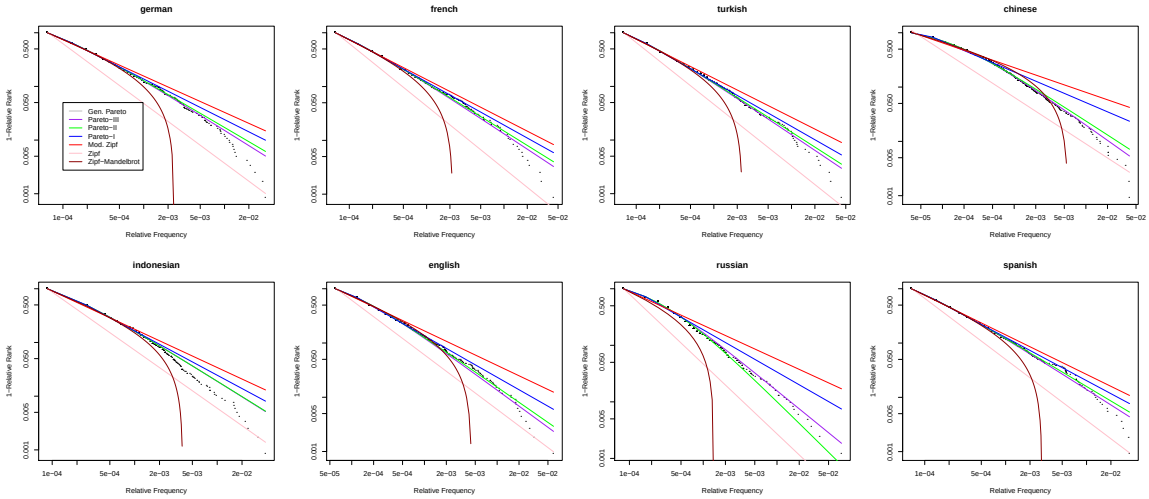

Figure 11: Log-Log CDF plot of word relative frequency versus inverse relative rank for "The little Prince".

Language Families

afro-asiatic	**chibchan**	danish	lukpa
hebrew	cabecar	spanish	
amharic		italian	**nilo-saharan**
coptic	**constructed**	french	dinka
tachelhit	esperanto	nepali	zarma
syriac		afrikaans	
wolaytta	**creole**	german	**oto-manguean**
tuareg	aukan	portuguese	amuzgo
somali	creole	swedish	chinantec
kabyle		norwegian	
arabic	**dravidian**	manx	**quechuan**
	telugu	english	quichua
algic	malayalam	gujarati	
potawatomi	kannada	romani	**sino-tibetan**
ojibwa			chinese
	equatorial	**iroquoian**	myanmar
altaic	camsa	cherokee	paite
turkish			
korean	**indo-european**	**japonic**	**tai-kadai**
	farsi	japanese	thai
arawakan	latvian		
campa	ukranian	**jivaroan**	**tucanoan**
	armenian	shuar	barasana
austro-asiatic	lithuanian	aguaruna	
vietnamese	croatian	achuar	**uralic**
	latin		hungarian
austronesian	hindi	**mayan**	finnish
chamorro	albanian	kiche	estonian
indonesian	polish	qeqchi	
uma	czech	uspanteco	**uto-aztecan**
malagasy	slovene	cakchiquel	nahuatl
cebuano	russian	jakalteko	
tagalog	bulgarian	mam	**west-papuan**
maori	greek		galela
	slovak	**niger-congo**	
basque	romanian	zulu	
basque	icelandic	xhosa	
	serbian	ewe	
carib	marathi	swahili	
akawaio	gaelic	wolof	

Table 4: Grouping of all analysed languages into their respective families

	Pareto III		Pareto II		Pareto I		Zipf (mod.)		Log-Normal		Burr		Log-Cauchy		Zipf (org.)		Zipf-Mandelbrot	
	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS	Sq. Error	KS
100	0.000022	0.004100	0.000004	0.000592	0.000005	0.002043	0.002396	0.024533	0.040302	0.003188	0.205219	0.038105	0.039486	0.000127	0.790015	0.249888	0.779849	0.248997
500	0.043866	0.060563	0.034992	0.050022	0.047384	0.063213	0.081811	0.084213	1.112249	0.079027	2.130549	0.107465	2.028688	0.017787	4.663579	0.274449	3.215130	0.248954
1000	0.026140	0.037953	0.026867	0.037523	0.026919	0.036400	0.026939	0.036709	1.931369	0.053344	3.600039	0.076878	2.616131	0.019662	1.927380	0.110341	0.231910	0.098123
5000	0.106204	0.006427	1.598728	0.282884	1.241508	0.121458	8.156660	0.605744	0.137575	0.021195	0.285452	0.020260	0.251352	0.040519	20.697590	1.203869	6.503947	0.480730
10000	0.119651	0.008050	2.920860	0.549974	1.295437	0.127016	7.139636	0.653773	0.137968	0.007791	0.453746	0.028206	0.301588	0.030985	8.588758	0.913065	6.884853	0.606076
Average	0.059177	0.023419	0.916290	0.184199	0.522251	0.070026	3.081488	0.280994	0.671893	0.032909	1.335001	0.054183	1.047449	0.021816	7.333464	0.550322	3.523138	0.336576

Table 5: Error Measures for randomly generated texts of different lengths.