Random Sampling of the Zipf-Mandelbrot Distribution as a Representation of Vocabulary Growth

Martin Tunnicliffe[1] **and Gordon Hunter**

School of Computer Science and Mathematics, Kingston University,
Penrhyn Road, Kingston-on-Thames, Surrey KT1 2EE, United Kingdom.

Abstract: We develop a discrete model of type-token dynamics based on random type selection from the Zipf-Mandelbrot probability distribution, with a view to examining the relationships between the constants of Zipf's and Heaps' laws. Analysis of items randomly selected items from the Standardised Project Gutenberg Corpus (SPGC) reveal a significant low-frequency "droop" in the β-slope of the types vs. frequency distribution, inconsistent with the model when vocabulary is unlimited: when a finite vocabulary limit is imposed, optimal parameter selection allows the droop to be reproduced. We adjust the parameters of both the limited and unlimited vocabulary models to obtain optimal agreement with the vocabulary growth curves: the limited vocabulary model usually yields the best optimised agreement, but a sizeable minority of items are better represented by an unlimited vocabulary. While the optimised Zipf α indices correlate strongly with the corresponding values obtained directly from document statistics, the former are generally larger than the latter (though this this is partially explained by the distorting effect of large values of the Mandelbrot parameter m). The β indices optimised from the limited vocabulary model are also compared with their directly measured equivalents, showing significant positive correlation. The relationship between optimised α and β agrees plausibly with the well-known continuum model, though the degree of agreement depends on how β is defined. The experiments yield repeatable results from each of three 100-item samples, demonstrating the statistical significance of the experiments.

Key words: Type/token systems, corpus linguistics, Zipf's law, Heaps' law, model optimization.

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

1. Introduction

Type/token systems, in which discrete entities or *tokens* belong to categories or *types*, are ubiquitous throughout the natural and artificial worlds [1]. They include biological habitats, where tokens are the individual organisms and types are the species or genera to which they belong; the Malayan butterfly populations studied by Corbett and modelled statistically by Fisher [2,3] led to some early insights, and more recent work [4,5] has yielded precise estimates of the number of unknown species awaiting discovery. Several statistical generalities have been observed [6]: Heaps' (or Herdan's) law relates the respective numbers of types and tokens, while the two Zipf laws (referred to here as Zipf's "first" and "second" laws) govern the frequency and ranking of types. Traditionally, one or other Zipf law has been considered fundamental, with the remaining Zipf law and Heaps' law emerging as mathematical consequences [7]. Recent work however suggests that all these laws may be emergent properties of underlying complex systems, which do not always remain stable as those systems expand [8].

Considerable attention has been paid to written documents, where tokens are the word instances and types are the dictionary words [9, 10]. This has several practical applications, including author attribution [11] and optimised information retrieval [12]. While some workers have looked only at

[1] Corresponding author.
E-mail addresses M.J.Tunnicliffe@kingston.ac.uk (M. Tunnicliffe), G.Hunter@kingston.ac.uk (G. Hunter).

individual documents, others have studied of entire languages: Petersen et al. [13] observed a "cooling with expansion" whereby fewer new words emerge as vocabulary grows, a trend punctuated by periods of "heating" associated with political conflict. One of these authors, Perc [14], studied the evolution of English words and phrases over several centuries, revealing a 200-year period of self-organization leading to statistical stabilization around the year 1800.

Table 1: Randomly selected samples from Standardized Project Gutenberg Corpus used in this study, all of which can be found at https://zenodo.org/record/2422561 [15]; the filename for each item is `PG<ID Number>_tokens.txt`.

Sample 1 Item ID Numbers									
32369	3011	52872	41607	19028	10899	21629	6280	7374	48123
42564	11255	32723	53375	3094	17741	27705	42086	8479	28655
26550	5268	41154	11988	21496	38698	50030	37243	51854	35458
53764	10383	2376	55541	16699	48497	1979	56408	50831	37107
17253	43948	54483	14379	32242	43186	23152	6768	28491	43983
1863	19951	36228	54078	24393	11696	11727	24621	21810	11947
33856	37047	38905	21720	49570	41460	12273	32042	27441	26233
32730	43953	21888	21092	28189	48027	22156	9311	36757	33041
34825	43485	467	12984	7098	25944	17274	53688	51712	1558
46365	36477	16253	25283	8724	56759	10718	19707	590	16448
Sample 2 Item ID Numbers									
53138	35600	10543	26390	19353	43225	18429	6850	102	33222
46111	46529	56009	56152	2053	12431	13026	13368	43452	44844
15205	37181	19503	49344	39360	40262	14778	46623	48069	55230
1397	35885	6164	41827	46594	10611	36790	57036	18330	55116
19356	49204	14000	8673	10641	46107	6973	23022	37584	44112
15149	36279	43737	28115	17035	2702	56676	48249	26154	1671
25419	14586	41305	6663	47241	37739	9324	29838	35640	7065
16378	28938	32129	14944	6640	7064	6715	49257	39483	16298
34653	43736	40276	36540	14453	3321	13565	18030	11079	46971
53735	5343	37998	8213	51753	40668	19907	26933	11030	56605
Sample 3 Item ID Numbers									
47994	5401	48390	44804	38514	19298	39843	50918	39231	43788
48358	17668	5961	29829	17418	52135	26315	21684	56187	50397
12186	7359	49283	10762	2188	4014	18547	32428	6384	40032
16074	31528	38151	45197	56473	10483	23780	14028	23031	35259
27575	33156	19045	30609	41575	48015	38969	46428	39513	41697
26034	31464	39105	28067	11556	37614	27222	46710	33305	5874
2876	50360	11128	2943	46008	22158	44784	14874	53757	43956
46382	24073	474	47350	36858	53007	37315	11848	34925	21660
1483	37293	4590	36217	15994	40263	38196	11678	50184	26103
18335	41122	45150	39453	41050	28106	31627	49471	55718	2741

The current work focusses on documents, selected randomly from the Standardized Project Gutenberg Corpus (SPGC) [15]. Three independent samples are selected, each comprising 100 unique documents of 50,000 to 100,000 word tokens (see Table 1). These are processed under the assumption that a word "type" constitutes any unique sequence of letters, regardless of any common stem; for example "boy", "boys", "boy's" and "boys'" are all counted as separate types. Following the "traditionalists", we assume Zipf's first law to be a basic system property (i.e. tokens selected randomly from a Zipfian distribution) and thus formulate a model with two variants: one in which vocabulary (number of types) is free to grow indefinitely, the other in which it has an artificially imposed ceiling. We optimise both models to fit the profiles of our sample texts and compare the optimised parameters with their independently measured values. Finally, we draw our conclusions.

2. Background and Initial Observations

2.1 Heaps' Law

Heaps' law relates the number of tokens t to the corresponding number of types $v(t)$, which we shall call the "vocabulary". As the system "grows" (i.e. as the text is processed) tokens are selected randomly from an "at-risk" pool. Initially most of these are unique, but as time progresses an increasing number of existing types reappear. This slows the growth of vocabulary, and over many selections the empirical relationship

$$v(t) \propto t^{\lambda} \tag{1}$$

is generally found to emerge. This known as Heaps' or Herdan's law [16] and the parameter $\lambda \in [0,1]$ may be called the "Heaps index" (see Figure 1). (Appendix A defines the various power-law indices used in this table.) Common sense suggests that as the unused types are depleted, $v(t)$ must saturate and the law cease to apply. Although this *is* observed in ideogrammic languages like Chinese (where types are semantic characters with limited supply) it is not typically the case in grammatical languages like English or French, even for very long documents [17,18]. While all languages' vocabularies are ultimately finite, few texts ever come close to exhausting all available words: the 2020 Oxford English Dictionary (OED) lists 171,476 distinct words and a typical 20-year-old knows about 42,000 [19] (though these refer to lexical word stems, excluding inflexions and proper names). In contrast, James Joyce's *Ulysses* (SPGC item PG4300) has only 28,998 word-types and extrapolation using Heaps' law (Figure 2) shows that it would need to be 10 times longer to reach the present OED vocabulary. Since *Ulysses* was written over seven years, this "ultra-*Ulysses*" would likely have taken Joyce a further sixty years, during which diachronic drift (the emergence of new words and the obsolescence of old ones) would have ensured a continued supply of unused types [20]. In fairness, *Ulysses* is not the best example since it contains many independently coined neologisms: Figure 2 compares it with a more conventional novel, Edith Nesbit's *The Railway Children* (PG1874) with $\lambda = 0.574$ (cf. 0.732 for *Ulysses*). Extrapolation shows the text would need to be 300 times longer (100 times longer than *Ulysses*) to rival the OED vocabulary.

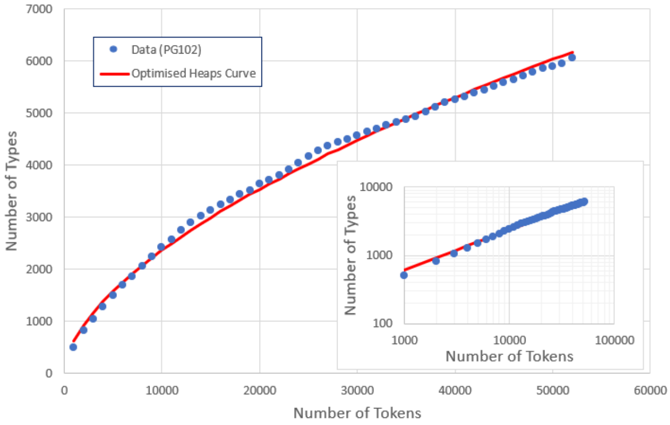


Figure 1: Vocabulary growth curve for SPGC item PG102 (*The Tragedy of Pudd'nhead Wilson* by Mark Twain) with optimized Heaps curve ($\lambda = 0.5818$) showing continuously slowing but non-saturating vocabulary growth. Inset shows the same graph plotted on a log-log scale. Optimization was performed by numerically minimising the sum of the square differences between the Heaps curve and the observed data.

2.2 Zipf's Laws

We define the *frequency* of a type as the number of tokens by which it is represented in a system. For example, "the" appears 3,635 times in the novel *Three Men in a Boat* and therefore has a frequency $f = 3635$. In its simplest form Zipf's law states the relationship between a type's frequency f_r and its *rank* r ($r = 1$ being the most frequent, 2 being the next most frequent etc.) as $f_r \propto \frac{1}{r}$ [21], although this has been generalized as the Zipf-Mandelbrot law:

$$f_r \propto \frac{1}{(r+m)^{\alpha}} \tag{2}$$

[22] where α may be called the "Zipf alpha index" and m the "Mandelbrot parameter" (which embraces an observed deviation from a power law when r is small). There is also the "Yule law" $f_r \propto \mathrm{B}(r, \alpha)$ [23] (where $\mathrm{B}(x, y) = \int_0^1 u^{x-1}(1-u)^{y-1}du$ is the Legendre beta function), which is asymptotically equivalent to $\frac{1}{x^y}$ and more tractably normalized. These are all variants of Zipf's "first law", but for the current paper we confine our definition of the law to (2).

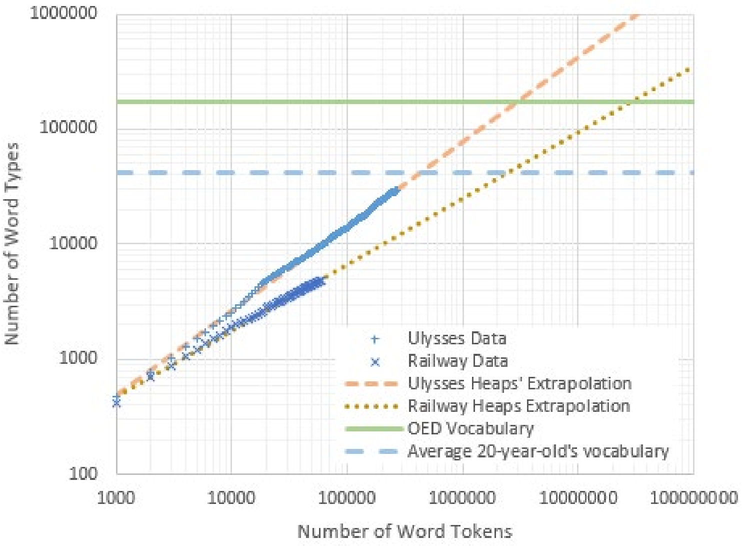
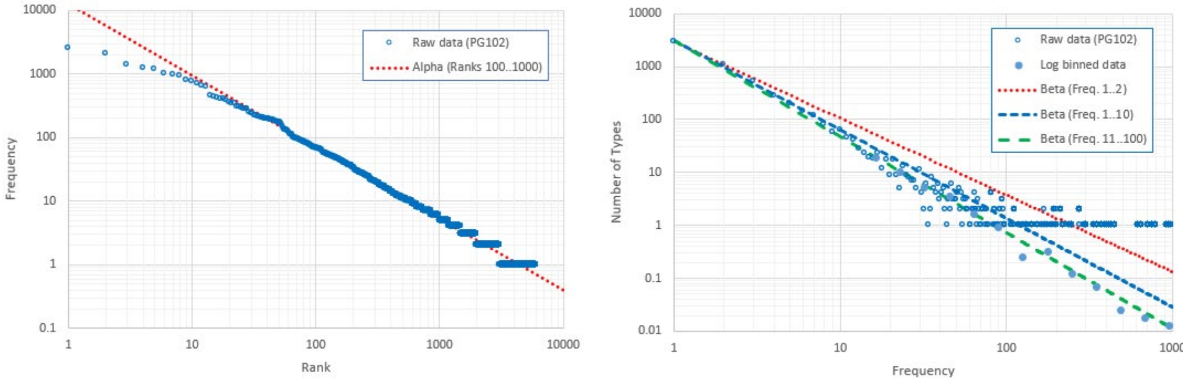


Figure 2: Vocabulary growth curves for *Ulysses* and *The Railway Children*, compared with the vocabularies of the Oxford English Dictionary (OED) and an average 20 year old. Heaps' extrapolation suggests that *Ulysses* would rival the OED after about 3,000,000 word tokens, while *The Railway Children* would require about 30,000,000.



(a) Frequency vs. rank

(b) Number of types vs. frequency

Figure 3: Frequency distributions for SPGC item PG102 (*The Tragedy of Pudd'nhead Wilson*) illustrating Zipf's first and second laws. (a) The frequency/rank curve is roughly continuous across the "mid range" (ranks 100..1000) with a slope (α) approximately 1.124, but becomes a series of discrete plateaux in the low frequency tail. (b) The Zipf β index becomes smaller when computed using lower frequencies; a general trend displayed in Figure 4. The discrete higher frequency data smoothed with logarithmic binning (using the scheme outlined in [24]) agree closely with the maximum likelihood estimation based on frequencies 11..100.

There is no universal consensus regarding the origin of Zipf's law. Whilst linguistically it could arise from optimal coding/decoding [22] or from the length and composition of words [25, 26], it applies also to article citations [27], human genetics [28], the sizes of galactic superclusters [29] and to a wealth of other non-linguistic phenomena [30]. Zipf himself linked his law to the principle of least effort [21], while others have proposed a guiding animus of "the rich getting richer" [31]. In the latter, type selections make future selections of those same types more probable – a mechanism also known as "preferential attachment" [32]. This sometimes incorporates a linear increase in the number of types-at-risk to prevent saturation [18], though Tria et al. [33] suggested that new-type selection itself expands the range of accessible types. Recent work by De Marzo et al. [34] showed that certain statistics including earthquake magnitudes and world city populations follow the rule only temporarily, and that Zipfian behaviour disappears as the sampling progresses. Davis [8] similarly proposed that Zipf's law may be a transitory phase during system growth, without any asymptotic convergence.

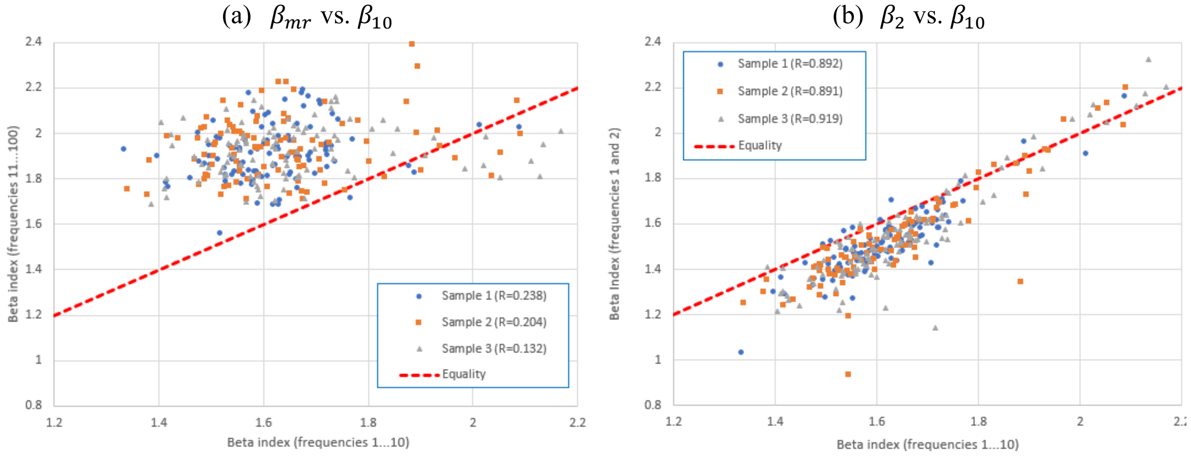
In textual documents we find that α, though supposedly constant, can vary quite considerably. A maximum likelihood estimate (MLE) of α in the "mid-range" $r = 100 \ldots 1000$ (ideally beyond the influence of m) is shown in Figure 3(a). (The MLE technique is outlined in Appendix B.) We denote this value α_{mr}, noting that it is not necessarily relevant to Heaps' law, which is governed by the low-frequency tail where new types are added. Although Montemurro [35] extended a smooth curve into this tail by averaging logarithmically spaced partitions, it is nevertheless better described using Zipf's "second law": that the number of types $n(f)$ exhibiting frequency $f = 1,2,3 \ldots$ is governed by

$$n(f) \propto \frac{1}{f^{\beta}} \qquad (3)$$

where β is the Zipf "beta index" (see Figure 3(b)). Since β quantifies the rate at which the plateaux to the right of Figure 3(a) widen with increasing r, it must vary inversely with the corresponding value of α: many authors (Lü et al. [7], Li [36] and others) have derived the formula

$$\beta = 1 + \frac{1}{\alpha} \qquad (4)$$

from a continuous approximation of Zipf's first law valid asymptotically for large f [37]. For our purposes we define three measures of β: the "mid range" β_{mr} obtained by applying MLE to the frequencies between 11 and 100, together with β_{10} and β_2 similarly computed using the frequency ranges 1 ... 10 and 1 ... 2 respectively. We find that (with the exception of a few outliers) $\beta_{mr} > \beta_{10} > \beta_2$ (Figure 4) which is consistent with the "droop" in the types vs. frequency distribution and the steepening of the frequency vs. rank curve observed by Montemurro [35], Tria et al. [33] and Cancho & Solé [38] for $r \gtrsim 10^4$.


(a) β_{mr} vs. β_{10}

(b) β_2 vs. β_{10}

Figure 4: Zipf beta indices obtained from the SPGC items listed in Table 1 using MLE. (a) The "mid-range" beta index β_{mr} computed from frequencies 11 to 100 plotted against the corresponding values of β_{10} based on the ten lowest frequencies, showing that the former is almost universally larger than the latter. There is a small and marginally significant positive correlation. (b) The index $\beta_2 = \log_2\big(n(1)/n(2)\big)$ based on the two lowest frequencies plotted against β_{10}, showing strong positive correlation with the former is almost universally lower than the latter. The results indicate a droop in the increasing number of types with decreasing frequency (see also Figure 3).

Figure 5(a) shows β_{10} plotted against α_{mr}, which indicates no significant correlation with most data-points clustered well below the theoretical prediction. (This confirms more robustly an earlier finding by the authors [39].) However, Figure 5(b) shows that β_{mr} and α_{mr} *are* related in a manner consistent with established theory.

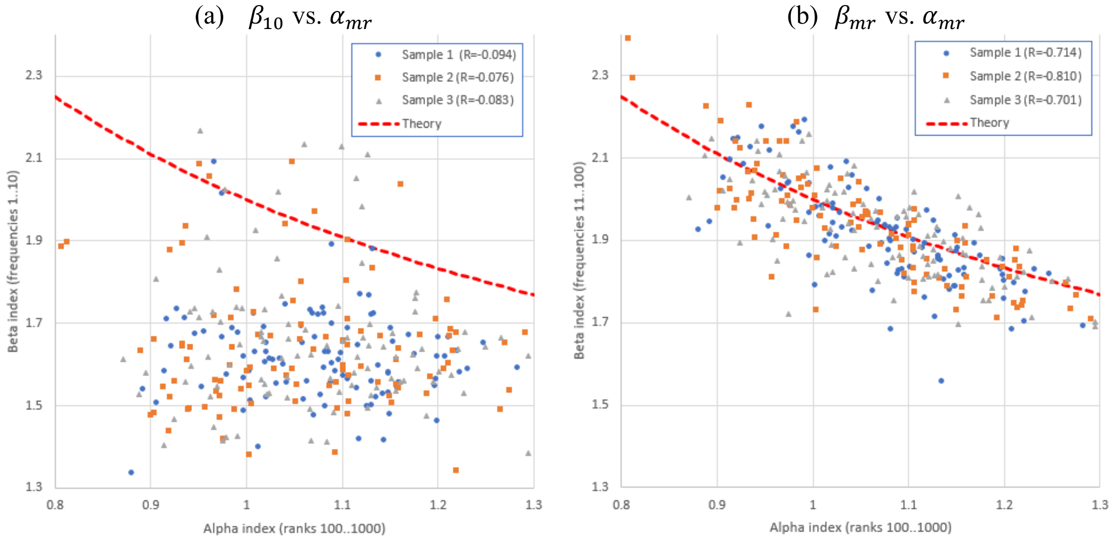
The relationship between Zipf's and Heaps' laws has been studied by many researchers, most of whom assume Bernoulli token selection with probabilities governed by a Zipfian distribution, to which relative frequencies tend asymptotically as the system expands. Based on this assumption, Boystov [40], van Leijenhorst & van der Weide [41], Lü et al. [7] and many others agree that Zipf's law leads to $v(t) = O(t^{\lambda})$ with

$$\lambda = \frac{1}{\alpha}. \tag{5}$$

Figure 6 shows that the measured λ is universally much lower than the value obtained by substituting α_{mr} into (5), although there is significant correlation in the required direction. However, we noted before that the low-frequency tail is more relevant to vocabulary growth than the mid-range, so a low-frequency β (β_{10} or β_2) may be more useful. Substituting (5) into (4) and solving for λ we obtain

$$\lambda = \beta - 1 \tag{6}$$

and Figure 7 compares this with the measured data. With $\beta = \beta_{10}$ (Figure 7(a)) the correlation becomes much more significant, with the data clustered plausibly around the theoretical curve, though with a lower than expected slope. One might expect better results to be obtained using β_2 (since it relates to the lowest frequencies most relevant to vocabulary growth) but Figure 7(b) shows that this is not the case: the centroid of the data now moves to the left of the line, and the coefficients of correlation decrease. The latter may be partly due to the fact that β_2 is based on fewer data than β_{10} and is therefore more susceptible to noise.


(a) β_{10} vs. α_{mr}

(b) β_{mr} vs. α_{mr}

Figure 5: Zipf β indices plotted against the corresponding mid-range α_{mr} (ranks 100 to 1000) indices for the SPGC items listed in Table 1, compared with the theoretical $\beta = 1 + 1/\alpha$ (4). All indices were obtained using the MLE method (Appendix B). Graph (a) shows β_{10} based on the ten lowest frequencies, while (b) shows β_{mr} based on frequencies 11 to 100. The former are mostly considerably lower than the model, with no significant correlation (Pearson R coefficients fail universally to meet the criterion for $p = 0.1$). The latter are clustered around the model curve, with very strong negative correlation.

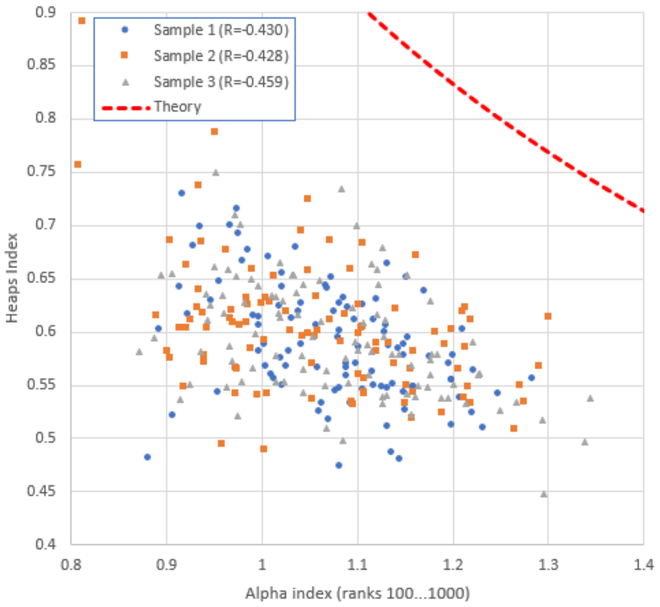
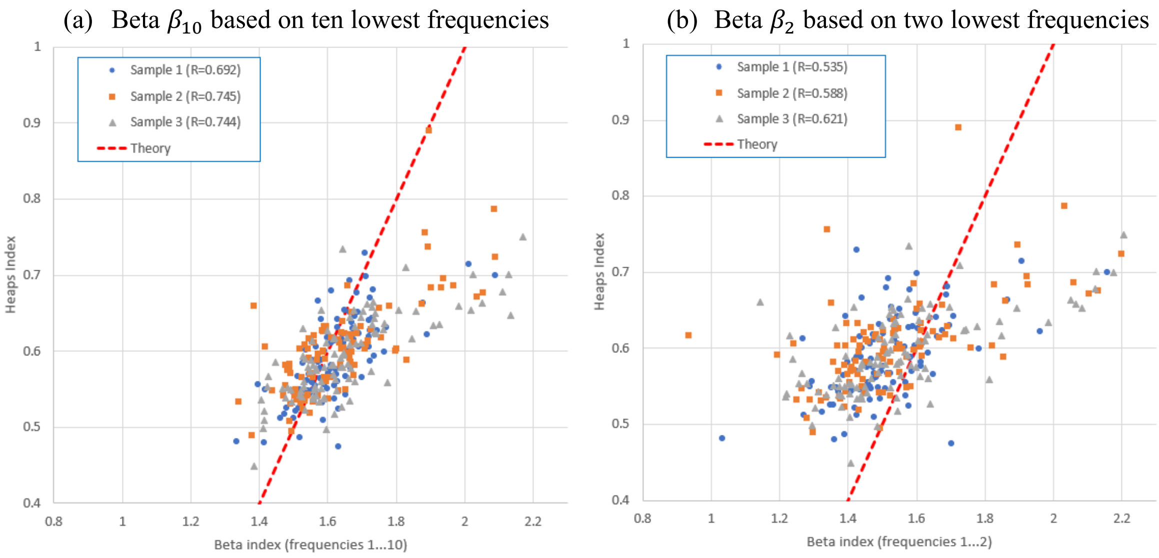


Figure 6: The optimised Heaps index λ plotted against the mid-range α_{mr}, with the curve of (5) shown for reference. The data are universally lower than the theoretical curve, though the strong negative correlation ($p \approx 0.00001$) shows that α_{mr} does have a significant impact upon λ.


(a) Beta β_{10} based on ten lowest frequencies

(b) Beta β_2 based on two lowest frequencies

Figure 7: Comparison of the observed relationship between λ and β with the theoretical $\lambda = \beta - 1$. The β-values were based on the slope of $n(f)$ vs. f for (a) the ten lowest frequencies (β_{10}) and (b) the two lowest frequencies (β_2). Note that in (b) there is a general bias towards over-prediction and a significantly weaker correlation than in (a).

Throughout the remainder of this paper we explore what use can still be made of traditional assumption of random selection from a static Zipfian distribution, using a combination of mathematical analysis and the SPGC documents of Table 1. We proceed on the assumption used by van Liejonhorst & van der Weide [41] that the Zipf-Mandelbrot law (2) is the most fundamental, and investigate the other laws in relation to it. We begin by examining the link between the Zipf indices α and β, before considering these in relation to the Heaps index λ.

3. Model Development

3.1 Relationship between the Two Zipf Laws

Most existing theory (e.g. Lü et al. [7]) relies on an approximation in which rank and frequency are represented as continuous numbers: $f(r) = \eta(r+m)^{-\alpha}$ where is η is a normalizing constant. If δr is the number of ranks associated with a small frequency change δf then (assuming $\delta r \ll r$) $\delta f \approx \eta[(r+m)^{-\alpha} - (r+m+\delta r)^{-\alpha}] \approx \eta\alpha(r+m)^{-\alpha-1}\delta r$. Substituting $r+m = (\eta/f)^{1/\alpha}$ and rearranging we obtain $\delta r \approx \frac{1}{\alpha}\eta^{\frac{1}{\alpha}} f^{-\left(1+\frac{1}{\alpha}\right)}\delta f$. Since the number of types $n(f)$ associated with a single frequency f is roughly the δr associated with $\delta f = 1$, it seems reasonable to state that

$$n(f) \approx \frac{1}{\alpha}\eta^{\frac{1}{\alpha}} f^{-\left(1+\frac{1}{\alpha}\right)} \tag{7}$$

which clearly implies (4). However, it is counterintuitive to use a continuous approximation to represent the *least* continuous part of the frequency vs. rank distribution. We therefore develop a discrete alternative based on the following assumptions: each token added to the document $C_t = \{S_1, S_2, \dots, S_t\}$ is selected from a "dictionary" $D_V = \{w_1, w_2, \dots, w_V\}$ of size V with an independent probability $p_r = \Pr[S_i = w_r]$, $i \in \{1,2,\dots,t\}$. (Note that t begins at 0 for an empty document and increases by 1 whenever a token is added.) We further assume Mandelbrot's version of Zipf's first law (see Section 2) governed by the distribution:

$$p_r = \frac{1}{\zeta_V(\alpha,m)(r+m)^{\alpha}}; 1 \le r \le V \tag{8}$$

where m is the Mandelbrot parameter and $\zeta_V(\alpha,m) = \sum_{i=1}^{V}\frac{1}{(i+m)^{\alpha}}$. (Appendix C outlines the procedure used to calculate this function.) We can call r the "rank" of the corresponding type, while noting that this is not necessarily its rank within the actual document.

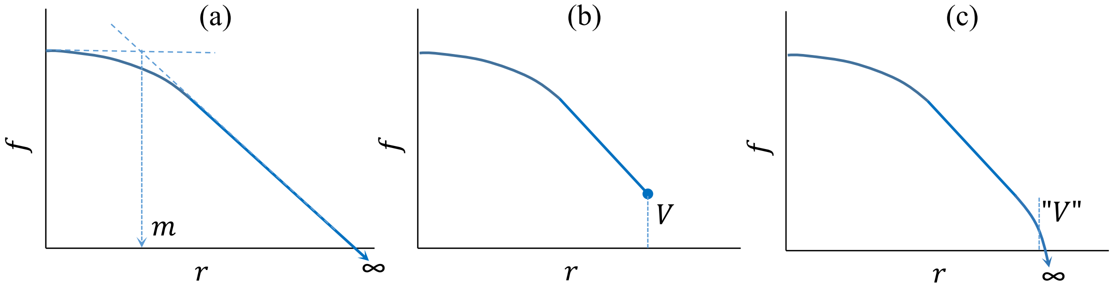


Figure 8: Conceptual probability/rank distributions: (a) Zipf-Mandelbrot law with unlimited vocabulary, (b) Zipf-Mandelbrot with artificially imposed finite maximum vocabulary V, (c) increased slope at low frequency observed by Montemurro [34],Tria et al. [32] and Cancho & Solé [42] represented by an "effective" maximum vocabulary V. (All scales are logarithmic.)

The maximum vocabulary V could be infinite as in Figure 8(a) or finite as in Figure 8(b), though in the former case normalization requires $\alpha > 1$. That a single Zipf-Mandelbrot distribution should remain undisturbed up to some precisely defined cut-off V may seem unnatural, especially in the light of the frequently observed gradual droop in the extreme low-frequency tail [33, 35, 38]: nevertheless, the assumption of a finite maximum vocabulary may help to mimic its effects for the purpose of modelling (see Figure 8(c)). (It should be noted that this low-frequency droop is somewhat speculative anyway, since it is only clearly observed in the aggregate of many different documents; in the case of Tria et al. [33] the entire Project Gutenberg corpus.)

For the extreme low-frequency tokens ($p_r \ll 1$) the Poisson approximation may be applied, so the probability that type w_r appears f times in document C_t is given by

$$p_t(f|w_r) \approx \frac{(tp_r)^f e^{-tp_r}}{f!} \tag{9}$$

and the number of types appearing exactly f times is

$$n_t(f) = \sum_{r=1}^{V} p_t(f|w_r) \approx \frac{t^f}{f!}\int_1^V {p_r}^f e^{-tp_r} dr. \tag{10}$$

Changing the variable of integration to p_r we obtain

$$\begin{aligned} n_t(f) &= \frac{t^f}{f!\,\alpha A^{\frac{1}{\alpha}}}\int_{\frac{1}{A(V+m+1)^\alpha}}^{\frac{1}{A(m+1)^\alpha}} {p_r}^{f-\frac{1}{\alpha}-1}\, e^{-tp_r} dp_r \\ &= \frac{1}{\alpha f!}\left(\frac{t}{A}\right)^{\frac{1}{\alpha}}\left[\Gamma\left(f-\frac{1}{\alpha},\frac{t}{A(V+m+1)^\alpha}\right)-\Gamma\left(f-\frac{1}{\alpha},\frac{t}{A(m+1)^\alpha}\right)\right] \end{aligned} \tag{11}$$

where $A = \zeta_V(\alpha, m)$ and $\Gamma(s,x) = \int_x^\infty u^{s-1}e^{-u}du$ is the upper incomplete gamma function [42]. Since $V \gg 1$, the $+1$ in the first term is negligible and if $m \ll V$ the final term may also be ignored. Thus

$$n_t(f) \approx \frac{1}{\alpha f!}\left(\frac{t}{\zeta_V(\alpha,m)}\right)^{\frac{1}{\alpha}}\Gamma\left(f-\frac{1}{\alpha},\frac{t}{\zeta_V(\alpha,m)(V+m)^\alpha}\right) \tag{12}$$

which for the case of unlimited types ($V \to \infty$) becomes

$$n_t(f) \approx \frac{1}{\alpha f!}\left(\frac{t}{\zeta(\alpha,m)}\right)^{\frac{1}{\alpha}}\Gamma\left(f-\frac{1}{\alpha}\right) \tag{13}$$

where $\Gamma(s) = \int_0^\infty u^{s-1}e^{-u}du$, the complete gamma function, and $\zeta(\alpha,m) = \lim_{V\to\infty}\zeta_V(\alpha,m)$ which equals the Riemann zeta function $\zeta(\alpha)$ when $m = 0$. (A similar expression to (13) was derived independently by Eliazar [43] on the basis of Poissonian token arrival.) Now for integer z, $z! = \Gamma(z+1)$ such that $\Gamma(z+1) = z\Gamma(z)$, which generalizes to $\Gamma(z+\gamma) = z^\gamma\Gamma(z)\left[1+\frac{1}{2z}\gamma(\gamma-1)+\mathcal{O}(|z|^{-2})\right], \gamma \in \mathbb{R}$ [44]. Substituting this into (13) we obtain

$$n_t(f) = \frac{1}{\alpha}\left(\frac{t}{\zeta(\alpha,m)}\right)^{\frac{1}{\alpha}} f^{-\left(1+\frac{1}{\alpha}\right)}\left[1+\frac{1+\alpha}{2\alpha^2 f}+\mathcal{O}\left(\frac{1}{f^2}\right)\right]. \tag{14}$$

Now setting $\eta = \frac{t}{\zeta(\alpha,m)}$ we find that the continuum approximation (7) is identical to (14) with all but the first bracketed term ignored. Therefore (given our assumptions) Zipf's first and second laws are only strictly consistent when $f \gg \frac{1+\alpha}{2\alpha^2}$, so it is hardly surprising that (4) does not work well in the low-frequency tail (Figure 5(a)). However, returning to our previous definition $\beta_2 = \log_2 \frac{n_t(1)}{n_t(2)}$ and substituting (13), we obtain after some manipulation

$$\beta_2 = 1 - \log_2\left(1 - \frac{1}{\alpha}\right) \tag{15}$$

which provides a discrete equivalent to (4) for extreme low frequencies under unlimited vocabulary (though (4) still holds asymptotically for larger frequencies). Interestingly, for all $\alpha > 1$ (15) returns a value larger than (4), conflicting with our previous observation that the extreme low frequency β is almost universally *smaller* than the average (Figure 4).

For a finite maximum vocabulary, we obtain a similar expression to (15) by substituting (12) in place of (13): with the aid of the identity $\Gamma(s+1,x) = s\Gamma(s,x) + x^s e^{-x}$ [42] we find

$$\beta_2 = 1 - \log_2\left(1 - \frac{1}{\alpha} + \frac{(p_V t)^{1-\frac{1}{\alpha}} e^{-p_V t}}{\Gamma\left(1 - \frac{1}{\alpha}, p_V t\right)}\right) \tag{16}$$

where $p_V = \frac{1}{\zeta_V(\alpha,m)(V+m)^\alpha}$. Since the final bracketed term is positive, this must always give a value lower than (15), suggesting the reduced values of β_2 seen in Figure 4 can be reproduced in the model by imposing an effective limit on the vocabulary.

3.2 Relationship Between Zipf and Heaps' Laws

Although this has been firmly established elsewhere, we note that (13) provides yet another proof of that Zipf's law leads to Heaps' law: theoretically $v(t) = \sum_{\forall f} n_t(f)$, but if we assume some upper frequency f_p beyond which further summation of $n_t(f)$ is negligible, (13) may be substituted for $n_t(f)$ to give

$$v(t) \approx \frac{1}{\alpha}\left(\frac{t}{\zeta(\alpha,m)}\right)^{\frac{1}{\alpha}} \sum_{f=1}^{f_p} \frac{\Gamma\left(f - \frac{1}{\alpha}\right)}{f!}. \tag{17}$$

Since the summation is independent of t we have $v(t) \propto t^\lambda$ with $\lambda = \frac{1}{\alpha}$ as required. However, before (17) can be used in any practical calculations, the issue of choosing f_p must be addressed: this cannot be arbitrarily large since the resulting high probabilities would invalidate the Poisson approximation (9). Nevertheless, as the frequency increases progressively fewer types exhibit the *same* frequency, until eventually a given frequency f_m can be associated with a single rank r_m with probability $p_{r_m} \approx \frac{f_m}{t}$. Substituting this expression into (8) and rearranging yields $r_m = \left(\frac{t}{\zeta(\alpha,m) f_m}\right)^{\frac{1}{\alpha}} - m$, which must equal the approximate shortfall in $v(t)$ obtained by summing terms in (17) up to the frequency $f_m - 1$. We can therefore state that

$$v(t) \approx \left(\frac{t}{\zeta(\alpha,m)}\right)^{\frac{1}{\alpha}} \left(\frac{1}{f_m{}^{\frac{1}{\alpha}}} + \frac{1}{\alpha} \sum_{f=1}^{f_m - 1} \frac{\Gamma\left(f - \frac{1}{\alpha}\right)}{f!}\right) - m. \tag{18}$$

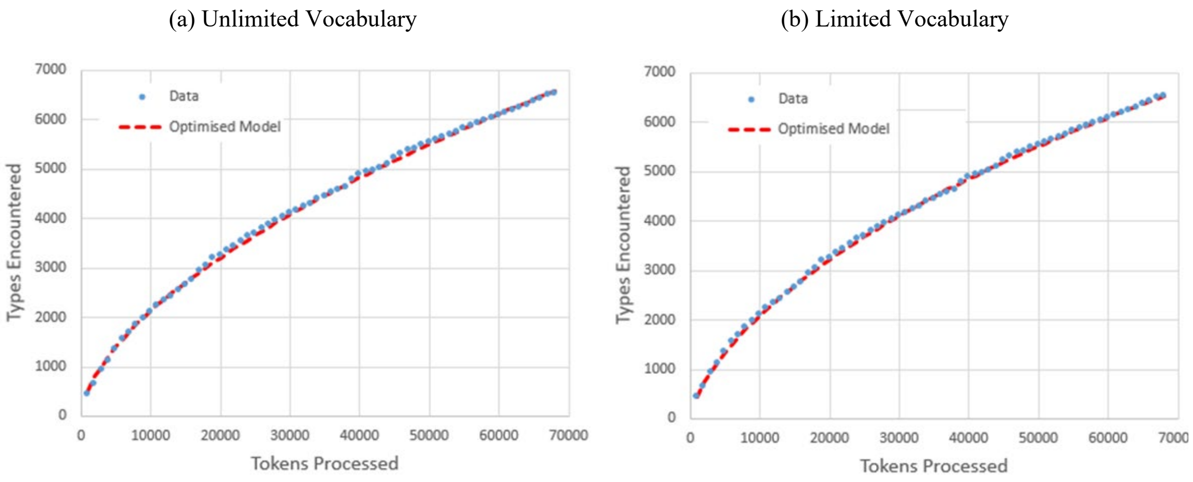
While the final term in (18) violates precise consistency with Heaps' law, if m is small relative to $v(t)$ the difference should only be slight. The choice of f_m is somewhat arbitrary; for our calculations we use the smallest f_m for which $n_t(f_m) < 0.1n_t(1)$.

The expression (18) assumes infinite maximum vocabulary; to allow V to be finite, we substitute (12) in place of (13) to obtain:

$$v(t) = \left(\frac{t}{\zeta_V(\alpha,m)}\right)^{\frac{1}{\alpha}} \left(\frac{1}{f_m^{\frac{1}{\alpha}}} + \frac{1}{\alpha}\sum_{f=1}^{f_m-1}\frac{\Gamma\left(f-\frac{1}{\alpha}, p_V t\right)}{f!}\right) - m \qquad (19)$$

where again $p_V = \frac{1}{\zeta_V(\alpha,m)(V+m)^{\alpha}}$. Not only does the presence of t within the summation further violate Heaps' law, but $v(t)$ must also saturate as r_m approaches V: a phenomenon never observed in practice. It may nevertheless provide a useful regional approximation within a limited range of t to accommodate the low frequency droop illustrated in Figure 8(c).

The approximations used in the development of this model are justified by simulation in Appendix E.


(a) Unlimited Vocabulary

(b) Limited Vocabulary

Figure 9: (a) Vocabulary growth curve for PG308 *Three Men in a Boat* compared with (18) using optimised parameters $\alpha = 1.732$, $m = 74.08$. The r.m.s. optimization error was 44.40 tokens. (a) The same growth curve compared with (19) using optimised parameters $\alpha = 1.396$, $m = 15.41$, $V = 19575.6$. The r.m.s. optimisation error here was 32.43 tokens.

4. Application of Model to Sample Texts

4.1 Analysis of an Individual Text

We begin by optimising (18) and (19) to fit the vocabulary growth profile of a single text: PG308 *Three Men in a Boat*. The optimisation algorithm (described more fully in Appendix D) evolves an initial hypothesis concerning the model parameters towards an optimal solution for which the root mean square (r.m.s.) error between the measured and theoretical $v(t)$ profiles is at a minimum. Figure 9(a) compares measured and optimised model $v(t)$ for the unlimited vocabulary model (18), and Figure 9(b) for the limited vocabulary model (19). The introduction of the finite vocabulary reduces the optimal error quite considerably (by 27%) which is to be expected given the extra degree of freedom.

Figure 10 shows the respective frequency vs. rank distributions obtained from the same optimised models ($f(r) = tp_r$ where p_r is computed from (8)) compared with the corresponding measured distribution. Figure 10(a) shows that for unlimited vocabulary considerable disagreement exists in the high frequency statistics: the introduction of a vocabulary limit (Figure 10(b)) significantly improves the fit, though the optimised slope is still somewhat greater across the mid-range than that of the measured distribution. Finally Figure 11 shows the types vs. frequency distributions, comparing optimised (12) and (13) with the measured data. We note that in the former case (Figure 11(a)) the predicted increase in β for very small frequencies is not reflected in the measured data, although the the introduction of the vocabulary limit (Figure 11(b)) does bring the theoretical and experimental distributions into better agreement.

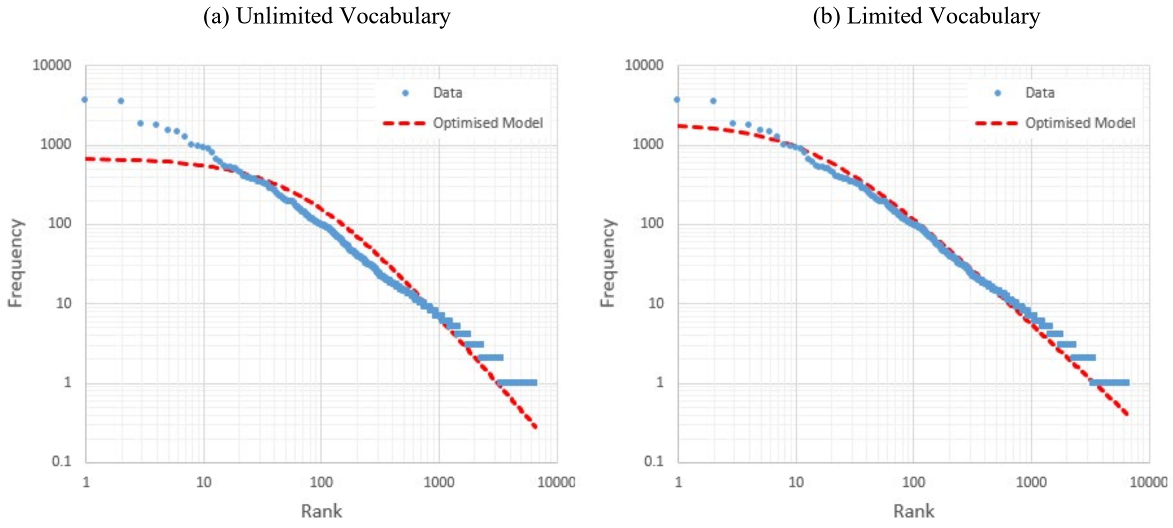


Figure 10: (a) Measured token frequency distribution for PG308 *Three Men in a Boat*, compared with *unlimited* vocabulary model ((8) with $V \to \infty$) using optimised parameters. (b) The same measured distribution compared with *limited* vocabulary model using optimised parameters.

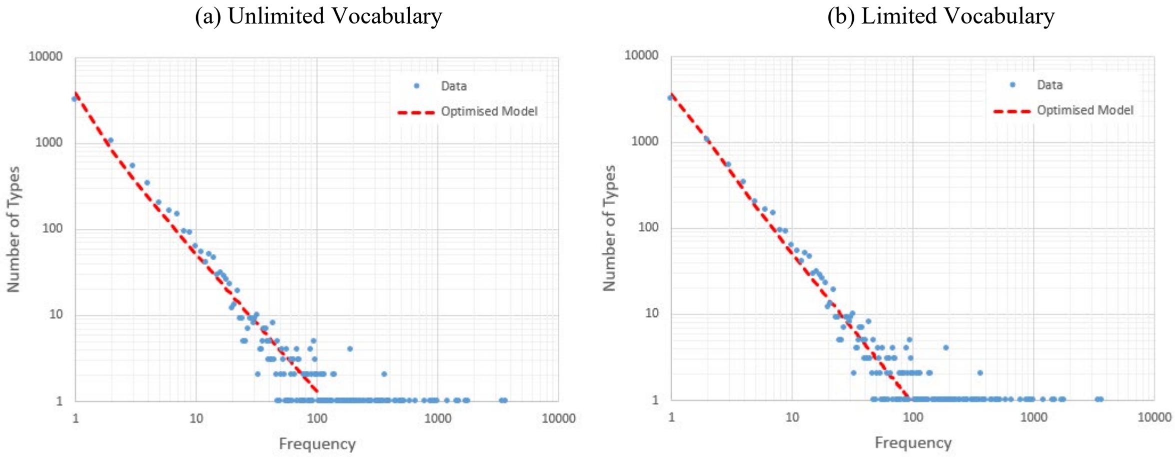


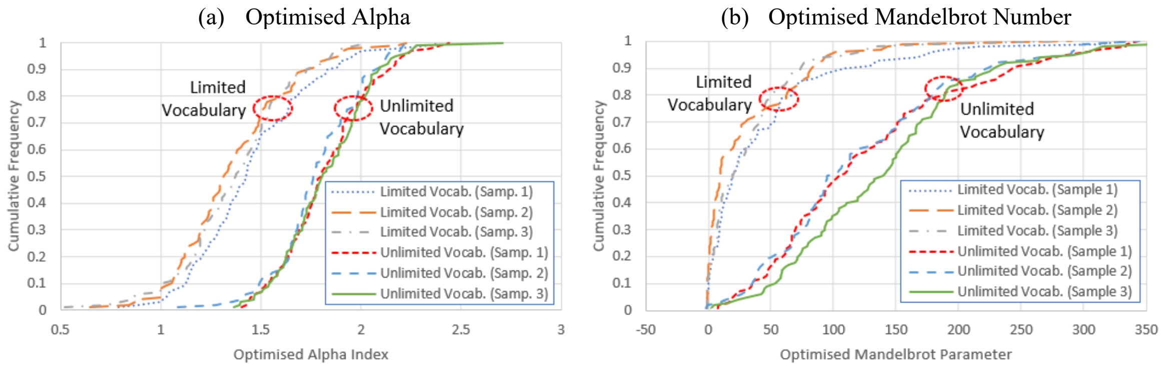
Figure 11: (a) Measured types vs. frequency distribution for PG308 *Three Men in a Boat*, compared with limited vocabulary model (13) using optimised parameters. (b) The same measured types vs. frequency distribution compared with (12) using optimised parameters.

4.2 Optimised Parameters for Project Gutenberg Samples

Figure 12 shows the cumulative distributions of optimised α and m for all three Gutenberg samples. The distributions of α are approximately Gaussian, with an average greatly reduced by the introduction of limited vocabulary (Figure 12(a)). There is a similar reduction in the average m, though the distribution is significantly different (Figure 12(b)): for unlimited vocabulary this is still basically Gaussian, while for limited V it resembles a gamma distribution with a shape parameter less than unity.

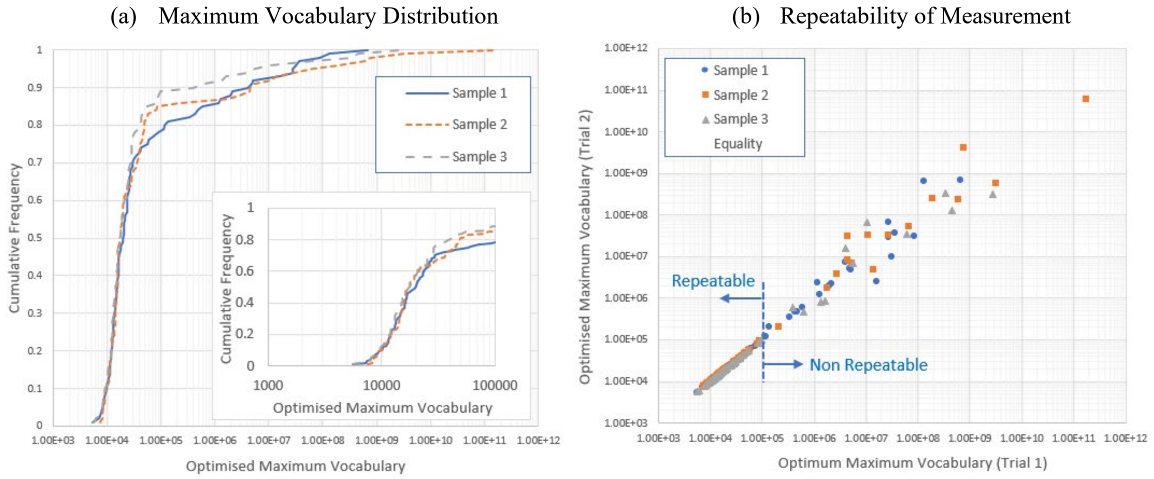
Figure 13(a) shows the cumulative distribution for optimised V: 70-90% of each sample remains lognormal, though with a heavy tail extending in (the case of sample 2) to more than 10^{11} types.

To ensure that these results are genuinely unique (rather than merely local minima), initial hypotheses are chosen at random and the optimisation procedure repeated for all items. With the unlimited vocabulary model, optimised parameters were almost identical (within ±1%) between repeated tests, as were the values of α and m obtained using the limited vocabulary model. However, the effect of repeated optimisation on V are shown in Figure 13(b): for $V < 10^5$ the results again differ by little more than a fraction of a percent, but this changes in the upper tail of the distribution where (especially for $V \gtrsim 10^6$) successive optimisations may be up to an order of magnitude apart.


(a) Optimised Alpha

(b) Optimised Mandelbrot Number

Figure 12: Cumulative frequency distributions of (a) optimised alpha indices and (b) optimised Mandelbrot parameters obtained from the Project Gutenberg samples (Table 1). Both are reduced significantly by the introduction of a vocabulary limit.


(a) Maximum Vocabulary Distribution

(b) Repeatability of Measurement

Figure 13: (a) Cumulative frequency distributions of optimised V. The inset shows an expansion of the steepest part of the distribution displaying approximately lognormal shape. (b) Comparison of optimised V between independent optimisations: for smaller V the experiment is accurately repeatable (within ±1%), while in the upper tail of the distribution ($V > 10^5$) measurements are not accurately repeatable.

Figure 14 shows that the improvement in the optimisation error caused by the introduction of limited V (observed anecdotally in Figure 9) is in fact almost universal. Only a very few items display a ratio $\varepsilon_V/\varepsilon_\infty > 1$, and for all of these $V > 10^5$ (which we have already observed is not consistently reproducible). Figure 15 shows that in all such cases the optimised values of α and m are practically identical for the limited and unlimited vocabulary models, and we conclude that here the unlimited vocabulary model provides the better description. We surmise that these are not in fact true optimisations associated with global (or even local) minima as shown in Figure 16(a): they lie instead upon a slope of ever-decreasing gradient, tending towards a “true” optimum at infinity (Figure 16(b)). The exit criteria are met arbitrarily in a manner dependent on the stochastic nature of the optimisation process itself, or on the randomly selected initial hypotheses. These values of V are therefore not particularly meaningful, except in so far as they are always very large.

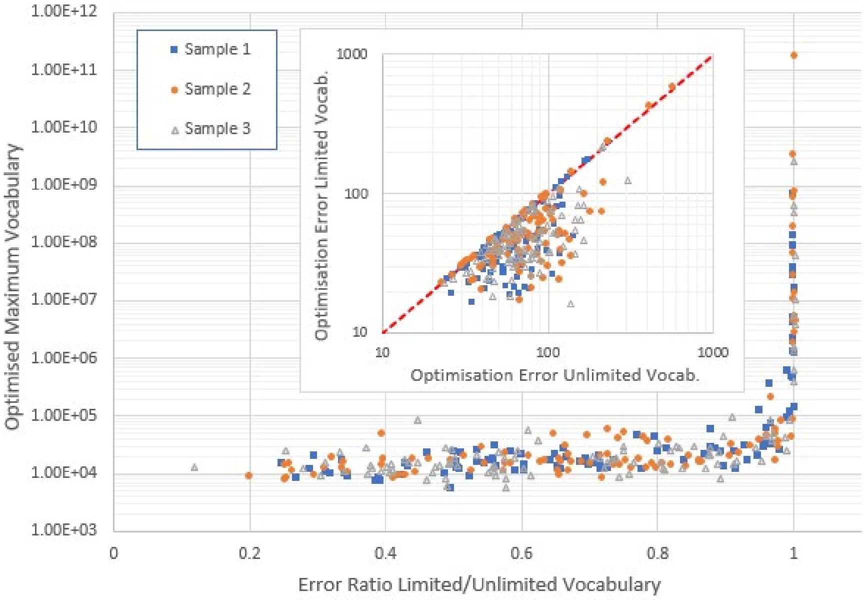
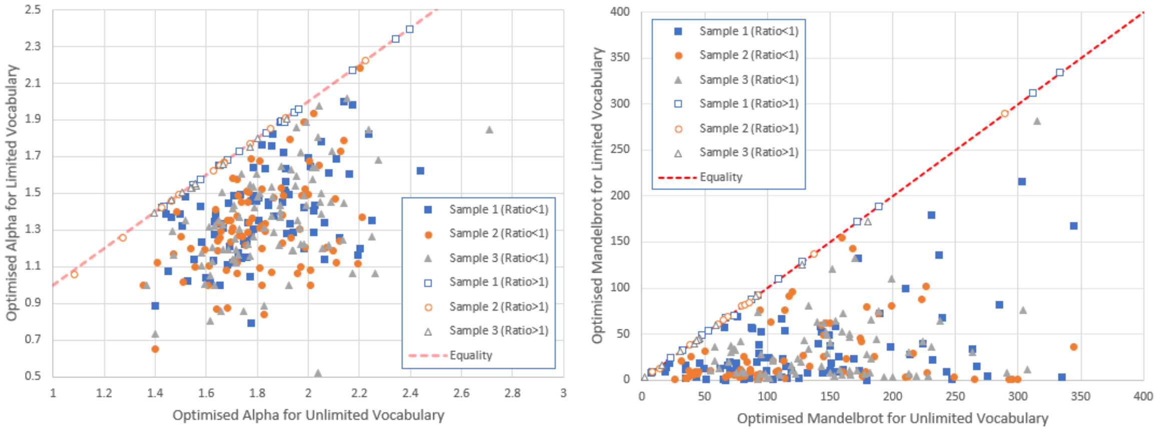


Figure 14: The impact on optimisation error of limiting V. The inset graph shows the limited V error ε_V plotted against the unlimited vocabulary error ε_∞, showing that the vocabulary limit generally improves accuracy. The main graph plots optimised V against the error ratio $\varepsilon_V/\varepsilon_\infty$ showing that the largest and therefore least repeatable (see Figure 13(b)) Vs appear when there is little or no accuracy improvement.



(a) Optimised α

(b) Optimised m

Figure 15: Comparison of optimised α and m values obtained using limited (19) and unlimited (18) vocabulary models. Note that when the ratio of the optimisation errors $\varepsilon_V/\varepsilon_\infty > 1$ (i.e. when the unlimited vocabulary gives the better fit) the parameters from the two models are practically identical.

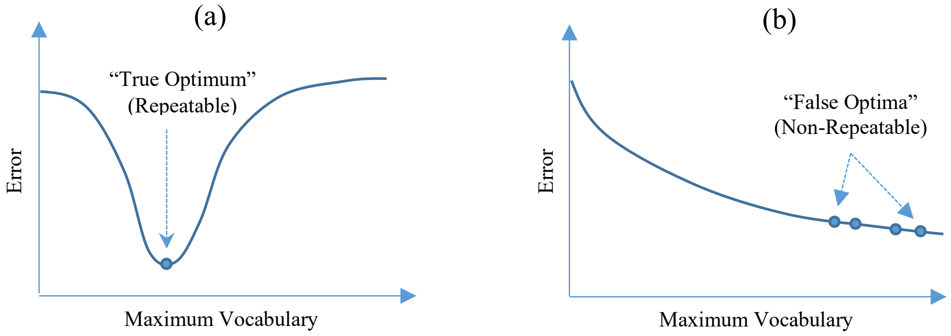


Figure 16: Schematic representation of (a) "true" parameter optimisation at a global (or local) minimum, and (b) "false" optimisation at arbitrary points along an ever-decreasing slope (the latter tending towards the true optimum at infinity).

4.3 Comparison of Optimised α and Mid-Range α_{mr}

Although the mid-range α_{mr} values were obtained by applying EM to the ranks 100-1000 (Figures 5 and 6), the optimised values of m were often well within this range (Figure 15(b)) such that measured α_{mr} depends in reality on both α and m. We therefore combine the optimised parameters to obtain a "reconstructed" $\bar{\alpha}_{mr}$ between ranks r_1 and r_2. By manipulating (8) we obtain

$$\bar{\alpha}_{mr} = -E\left[\frac{d\log p_r}{d\log r}\right] = \frac{1}{r_2 - r_1}\int_{r_1}^{r_2} \frac{\alpha r}{r+m}\,dr = \alpha\left(1 - \frac{m}{r_2 - r_1}\log\frac{r_2+m}{r_1+m}\right) \qquad (20)$$

for which we set $r_1 = 100$ and $r_2 = 1000$. Figures 17 and 18 compare the optimised values of α with the predictions of (20) obtained using the unlimited and limited vocabulary models, plotted against the directly measured α_{mr} for the same items. The mean "overestimations" $(\alpha - \alpha_{mr})$ and $(\bar{\alpha}_{mr} - \alpha_{mr})$ are plotted to the right of both graphs, indicating that (20) gives an improved estimation in both cases. However, there is still an overall positive bias, suggesting that notwithstanding the effects of m, the value of α relevant to vocabulary growth is appreciably larger than that of the mid-range.

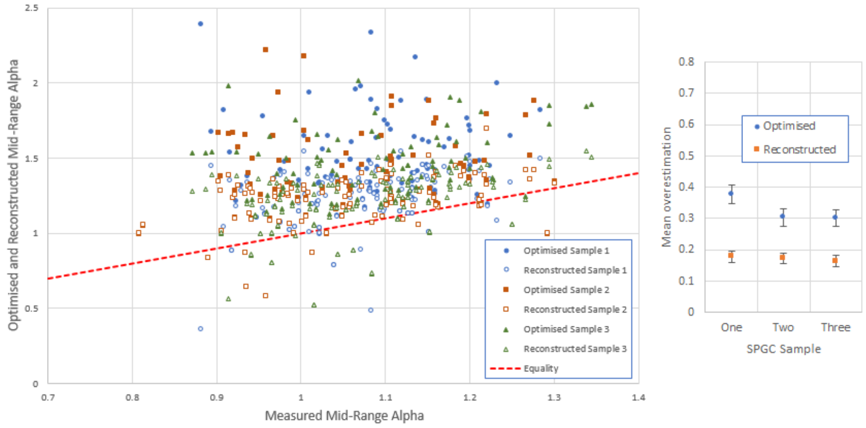


Figure 17: α_{mr} from frequency vs. rank distributions, compared with corresponding values optimised using the unlimited vocabulary model, and mid-range values $\bar{\alpha}_{mr}$ reconstructed using (20). The graph to the right shows the average overestimations $\alpha - \alpha_{mr}$ and $\bar{\alpha}_{mr} - \alpha_{mr}$, error bars indicating the 95% confidence interval for the mean ($\pm 1.98\,\sigma/\sqrt{n}$).

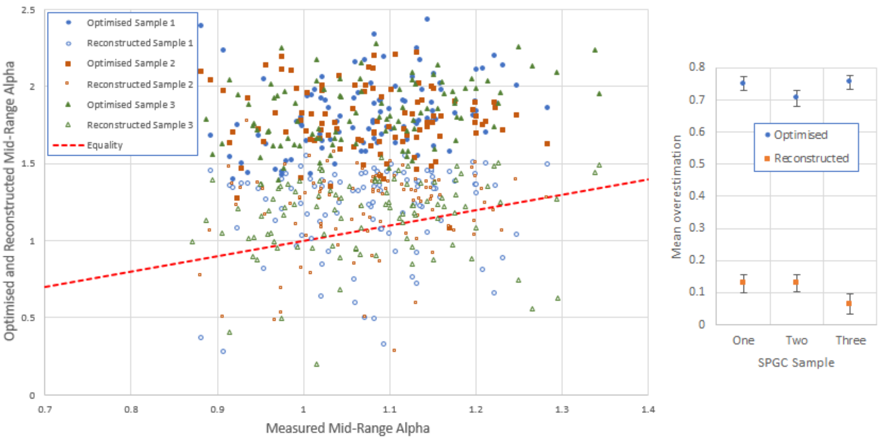
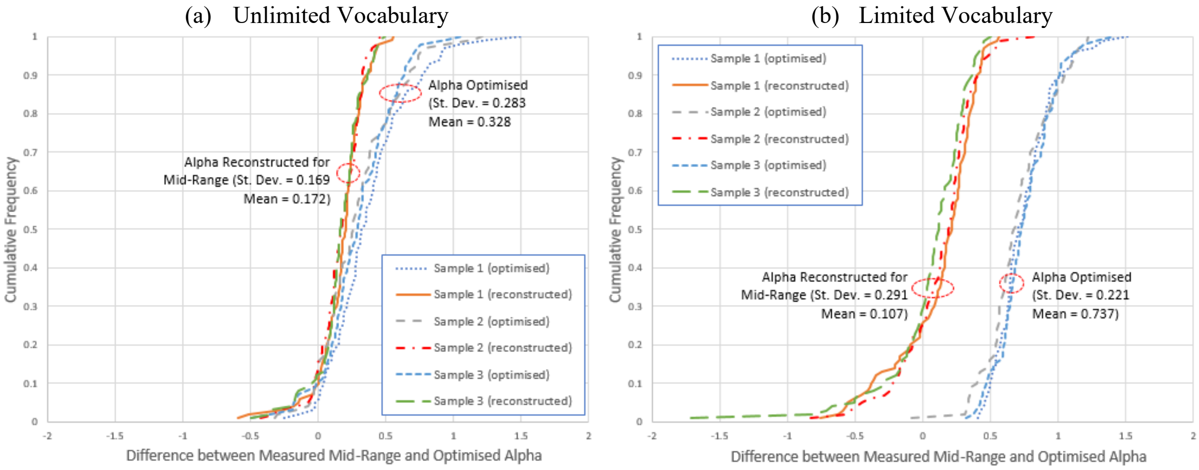


Figure 18: α_{mr} from frequency vs. rank distributions, compared with corresponding values optimised using the limited vocabulary model, and mid-range values $\bar{\alpha}_{mr}$ reconstructed using (20). The graph to the right shows the average overestimations $\alpha - \alpha_{mr}$ and $\bar{\alpha}_{mr} - \alpha_{mr}$, error bars indicating the 95% confidence interval for the mean ($\pm 1.98\,\sigma/\sqrt{n}$).

Figure 19 shows these same overestimations plotted as cumulative distributions. For the unlimited vocabulary model we note that the lower tails of the two distributions (where the overestimation is close to zero) are almost exactly convergent, indicating that here m has negligible impact. For the limited vocabulary model the distributions are more clearly separated. Interestingly, the $(\bar{\alpha}_{mr} - \alpha_{mr})$ values are almost twice as widely dispersed for the limited vocabulary model than for the unlimited vocabulary model, and a greater proportion of them (around 25-30%) are less than zero (indicating underestimation of the true α_{mr}). The close agreement between the distributions obtained from the three independent Project Gutenberg samples suggests that these effects are genuine, and are not artifacts of the limited sample sizes.


(a) Unlimited Vocabulary

(b) Limited Vocabulary

Figure 19: Alpha indices measured from frequency vs. rank distributions, compared with corresponding values reconstructed from (a) unlimited vocabulary and (b) limited vocabulary models optimised to fit the measured vocabulary growth curves. On average both models over-predict the measured value; by about 0.3 for the unlimited vocabulary model, and 0.2 for the limited vocabulary model.

4.4 Analysis of Measured and Optimised Zipf Indices

Figure 20 shows the measured β indices plotted against the optimised values of α obtained using the unlimited vocabulary model, along with the theoretical curves predicted by (4) and (15). We see that the discrete model (15) grossly overestimates both β_{10} and β_2, while the continuous model (4) skims the underside of the β_{10} data. Interestingly (4) *does* plausibly agree with the measured β_2 values, though the latter are extremely variable (each data point being based on only two measurements).

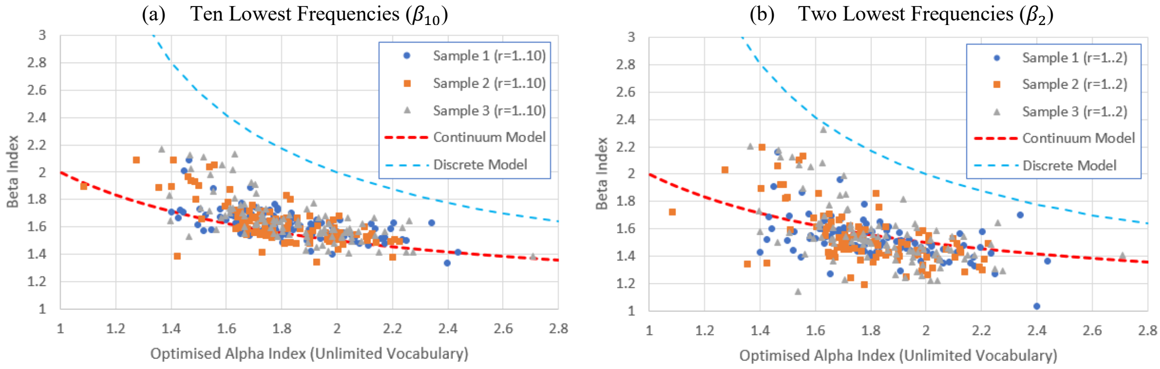


Figure 20: Relationship between the α indices optimised using the unlimited vocabulary model (18) and the β indices measured directly from the samples using (a) the ten lowest frequencies and (b) the two lowest frequencies. The broken lines indicate for reference the continuum model (4) and the discrete model for the two lowest frequencies (15).

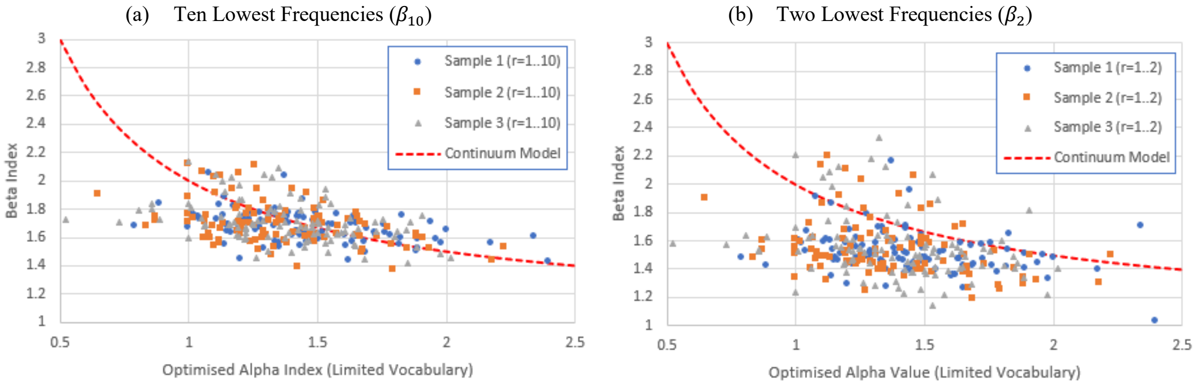


Figure 21: Relationship between the α indices optimised using the limited vocabulary model (19) and the β indices measured directly from the samples using (a) the ten lowest frequencies and (b) the two lowest frequencies. The broken line indicate for reference the continuum model (4).

Figure 21 shows the same analysis applied to the optimised values of α obtained using the limited vocabulary model (19), along with the continuum model curve predicted by (4). (The corresponding curve of the discrete model (16) cannot pe plotted on the same axes, since it requires the additional parameter V.) We now find that it is β_{10} which gives better agreement with the continuum model, while the latter skims the topside of the β_2 data. Figure 22 compares the measured values of β_2 with the corresponding predictions of the discrete model (16) using the optimised parameters α, m and V. We find both data-sets are statistically equivalent around 1.5, the regression line passing through this point. There is a weak but significant positive correlation, though the slope is almost an order of magnitude too small.

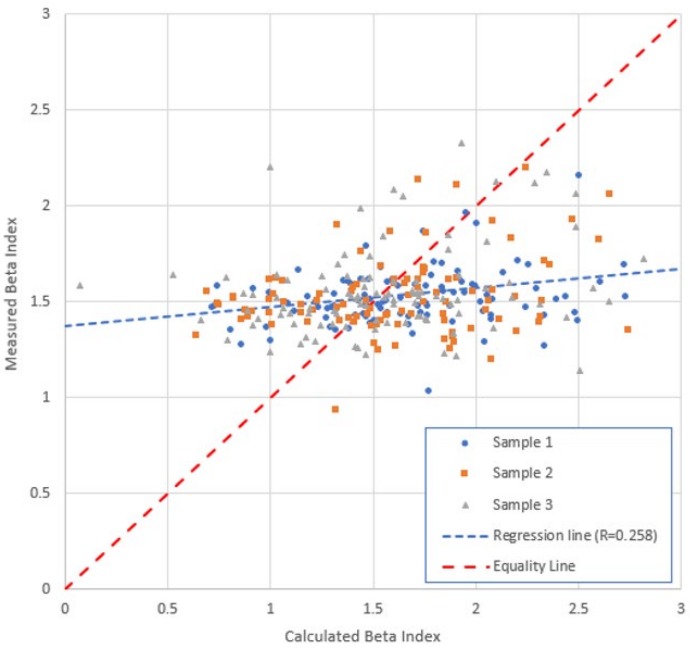


Figure 22: Relationship between the directly measured β_2 indices and those computed with (16) using the optimised parameters of the limited vocabulary model (19). The broken lines indicate the expected result (equality) and the regression line based on the aggregate data. The two lines agree approximately when $\beta_2 \approx 1.5$.

5. Conclusions and Future Work

In this paper we develop a discrete type-token model based on random selection from a Zipf-Mandelbrot probability distribution. Versions of this model, assuming unlimited and limited vocabularies, are optimised so as to agree with the observed vocabulary growth curves of a range of 50,000-100,000 word texts, selected at random from the Standardised Project Gutenberg Corpus (SPGC) and randomly grouped into three samples of 100 items each. The optimised parameters for each item were compared with values gleaned independently from the statistical distributions of those same texts. We make the following observations:

1. The data supposedly governed by Zipf's second law exhibit a low frequency "droop" such that the beta index for the two lowest frequencies β_2 (those most relevant to vocabulary growth) is generally lower than that obtained using MLE across wider frequency ranges. This hints at the low-frequency steepening of the frequency vs. rank distribution observed by Montemurro [35] and Tria et al. [33] and others.
2. When comparing the mid-range Zipf alpha (α_{mr}) and beta (β_{mr}) indices, the well-known equation $\beta = 1 + \frac{1}{\alpha}$ agrees with the general trend, though with a wide statistical scatter. However, the other widely-reported formula $\lambda = \frac{1}{\alpha}$ shows no agreement with the measured α_{mr} vs. Heaps' (λ) indices beyond a strong correlation in the required direction. The combined expression $\beta = 1 + \lambda$ agrees better when the beta index is estimated over the ten lowest frequencies (β_{10}) than over the two lowest (β_2), though in both cases the slope is less than the expected unity.
3. Assuming random selection from a Zipf-Mandelbrot distribution with no vocabulary limit, the number of types exhibiting a given frequency can be expressed in terms of the gamma function, with Zipf's second law valid asymptotically for large frequencies. Contrary to observation 1, the resulting β_2 is larger than its asymptotic value.

4. By summing gamma function expressions, we obtain a vocabulary growth function $v(t)$ almost exactly consistent with Heaps' law (barring an additive term equal to the Mandelbrot parameter m). This function can be optimised to agree quite closely with the measured vocabulary curves.
5. By introducing a maximum vocabulary limit, the model can be reformulated in terms of the incomplete gamma function, yielding modified expressions for β_2 and $v(t)$. The former contains a new term, which could potentially compensate for the increased β_2 mentioned in observation 3, and thus replicate the droop noted in observation 1.
6. In general, the limited vocabulary model gives better agreement with the measured data than does the unlimited vocabulary model. However, in the minority of cases where the optimised vocabulary V exceeds about 10^5 tokens, the latter ceases to be reproduced consistently between repeated optimisations. In these cases the unlimited vocabulary model fits the data best.
7. For both limited and unlimited vocabulary models, the average optimised α is significantly larger than α_{mr}, even when the distortion caused by the Mandelbrot parameter m is taken into account. The two α's are nevertheless strongly correlated. This adds further credence to the speculation appended to observation 1.
8. When α is optimised using the unlimited vocabulary model, the directly measured β_2 agrees plausibly with $\beta = 1 + \frac{1}{\alpha}$, while β_{10} is underpredicted. The situation is reversed when α is optimised using the limited vocabulary model: the equation agrees plausibly with β_{10}, while β_2 is overpredicted.
9. The values of β_2 computed from the limited vocabulary model agree somewhat with the directly measured values. The correlation is positive and significant, and the graphs of the two distributions cross at about $\beta_2 = 1.5$.
10. All experiments yielded statistically consistent results from each of the three 100-item samples, demonstrating the statistical significance repeatability of our experiments.

We note that even our most accurate model (random selection from Zipf-Mandelbrot distribution with a limited vocabulary) is still quite a blunt instrument; it uses an abrupt cut-off to represent what is almost certainly a continuous steepening of the frequency vs. ranks distribution and thus predicts a vocabulary saturation contrary to what is observed in practice (see Figure 2). Furthermore, the model is built upon a static probability distribution, while many researchers have found dynamically evolving processes better applicable to complex systems (e.g. Tria et al. [33]). Finally, the corpus items are selected at random with reference only to their lengths: though this is reasonable to eliminate selection bias, it means that texts exhibiting homogeneous vocabulary growth are placed together with ones displaying temporary bursts and lulls of new vocabulary. By screening out the latter, it may be possible to reduce the noise in the data and uncover clues not yet visible.

Appendix A. Definition of Power-Law Exponents

Papers on type-token theory use a range of conflicting names for the different power-laws, and symbols for their exponents. Table A.1 summarises the terminology used in this paper.

Table A.1: Power-law terminology

Name	Describes	Exponent
Zipf's first law	Rank vs. frequency	Alpha (α)
Zipf's second law	Frequency of frequency	Beta (β)
Heaps' Law	Vocabulary growth	Lambda (λ)

Appendix B. Maximum Likelihood Estimation of the Index in a Power Law Distribution

Historically many techniques have been used to estimate the index β for distributions of the form $p(x) \propto 1/x^{\beta}$, but maximum likelihood estimation (MLE) gives an unbiased result with minimal statistical variability [24, 45, 46]. If a variable X is power-law distributed across the range $A \leq X \leq B$, then the probability of outcome $X = x$ is given by

$$p(x|\beta) = \frac{1}{\zeta_{A,B}(\beta) x^{\beta}} \tag{B.1}$$

where $\zeta_{A,B}(\beta) = \sum_{i=A}^{B} \frac{1}{i^{\beta}}$. If the observed outcomes are $\{x_1 \ldots x_n\}$ then the likelihood of any particular value of β is $\prod_{i=1}^{n} p(i|\beta)$. It is easier to work with the logarithm of this expression, so we define $\mathcal{L}(\beta) = -n \log \zeta_{A,B}(\beta) - \frac{\beta}{n} \sum_{i=A}^{B} n(i) \log i$, where $n(i)$ is the number of observed outcomes equal to i. To find the value of β which maximises this function, we set $\mathcal{L}'(\beta) = 0$ and rearrange to obtain the implicit expression

$$\frac{\zeta'_{A,B}(\beta)}{\zeta_{A,B}(\beta)} = -\frac{1}{n} \sum_{i=A}^{B} n(i) \log i \tag{B.2}$$

where $\zeta'_{A,B}(\beta) = -\sum_{i=A}^{B} \frac{\log i}{i^{\beta}}$. Although approximate analytic solutions to (B.2) have been devised [46], β can easily be obtained numerically to any required degree of accuracy. Note that if we consider only the two lowest data points ($A = 1, B = 2$) the solution becomes $\beta = \log_2 \frac{n(1)}{n(2)}$.

Appendix C. Computing the Zipf-Mandelbrot Normalization Constant

To aid computation we use the approximation $\zeta_V(\alpha, m) \approx \sum_{i=1}^{i_m} \frac{1}{(i+m)^{\alpha}} + \int_{i_m+1}^{V} \frac{dr}{(r+m)^{\alpha}} = \sum_{i=1}^{i_m} \frac{1}{(i+m)^{\alpha}} + \frac{1}{\alpha-1}\left[\frac{1}{(i_m+1+m)^{\alpha-1}} - \frac{1}{(V+m)^{\alpha-1}}\right]$ which, assuming $V > i_m$, allows V to be a non-integer and if necessary infinite. (In the latter case we use the notation $\zeta(\alpha, m)$.) For all our calculations we set $i_m = 5000$.

Appendix D. Optimization Algorithm

The following process applies to the finite vocabulary model (19): the infinite vocabulary version is identical, with only two parameters and is based on (18). Since the optimal Mandelbrot number m is sometimes very close to zero (or even negative) we find it easier to redefine the model in terms of the parameter $\mu = m + 3$, adjusting the equations accordingly.

We define the optimization problem as the discovery of α, μ and V which minimize the error $\varepsilon(\alpha, \mu, V) = \sqrt{\sum_{i=1}^{\lfloor T/\Delta t \rfloor} \left[v_{\alpha,\mu,V}(i\Delta t) - \hat{v}(i\Delta t)\right]^2 / \lfloor T/\Delta t \rfloor}$ where $\hat{v}(t)$ is the measured vocabulary curve, $v_{\alpha,\mu,V}(t)$ is the model prediction, Δt is a suitably chosen sampling interval (we use 1000 tokens) and T is the final number of tokens in the corpus.

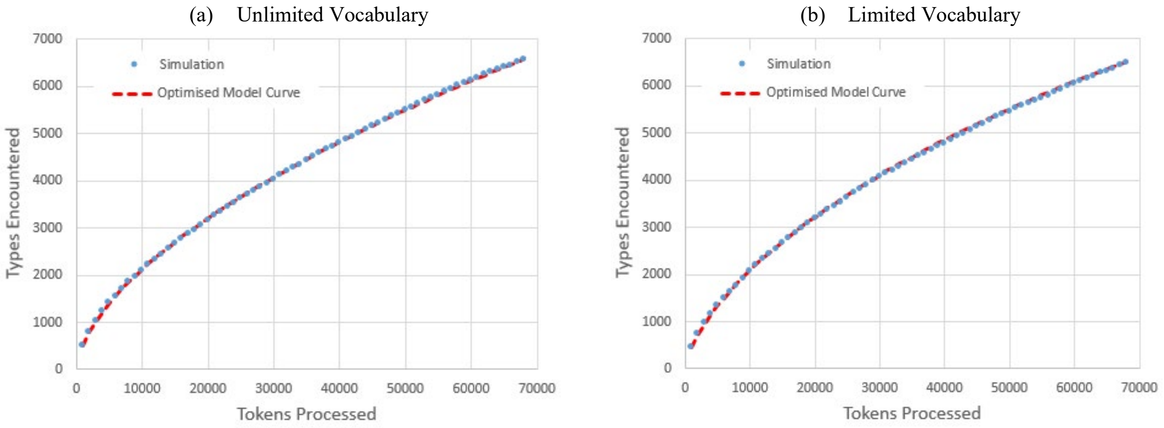
Optimization begins with a randomly selected initial hypothesis: α_1 between 1 and 2, m_1 between 0 and 150 (1.5 to 2.5 and 20 to 270 in the case of unlimited vocabulary) and V_1 between 10,000 and 100,000. The following iterative process is then applied. (The cycle number j is initially set to 1 and the step size and step-size S_1 to 10^{-5}).

1. Establish the function $F_j(x, y, z) = \varepsilon\big(\alpha_j[1 + x], \mu_j[1 + y], V_j[1 + z]\big)$. (The arguments x, y and z are not yet specified.)
2. Set the "working step-size" $S_j{}' = S_j$.
3. Set the non-beneficial mutation counter $i = 0$.
4. Create a "random mutation" by selecting independent Gaussian random variables x, y and z, each with a standard deviation $S_j{}'$ and zero mean. To prevent extreme outliers from derailing the procedure, outcomes are constrained to the range $(-1,1)$.
5. Compute $F_j(x, y, z)$.
6. If $F_j(x, y, z) < F_j(0,0,0)$ then accept the mutation as "beneficial" and go to step 9. Otherwise increment i.
7. If $i = 100$ then reduce $S_j{}'$ by 10% and reset $i = 0$.
8. Go to step 4.
9. Compute the magnitude of the mutation $d = \sqrt{x^2 + y^2 + z^2}$ and the corresponding gradient $\left|\nabla F_j\right|$ computed across finite intervals $\delta x = \delta y = \delta z = 10^{-9}$.
10. Update the model $\alpha_{j+1} = \alpha_j(1 + x)$, $\mu_{j+1} = \mu_j(1 + y)$ and $V_{j+1} = V_j(1 + z)$.
11. If $\left|\nabla F_j\right| < 0.01$ or $S_j < 10^{-9}$ then go to step 14.
12. Update the step-size to a weighted average of itself and the magnitude of the most recent mutation: S_j=$0.9S_{j-1} + 0.1d$. (This communicates the need for a step-size adjustment.)
13. Increment j and go to step 1.
14. The exit condition is now met. Report α_{j+1}, μ_{j+1} and V_{j+1} as the optimal solution, along with the solution error $\varepsilon\big(\alpha_{j+1}, \mu_{j+1}, V_{j+1}\big)$.

Despite the random starting hypotheses, the final parameters are replicated almost exactly by repeated optimisations (with the exception of the larger values of V, see Figure 13(b)), providing evidence that the minima are global and that the optimisations are therefore genuine.

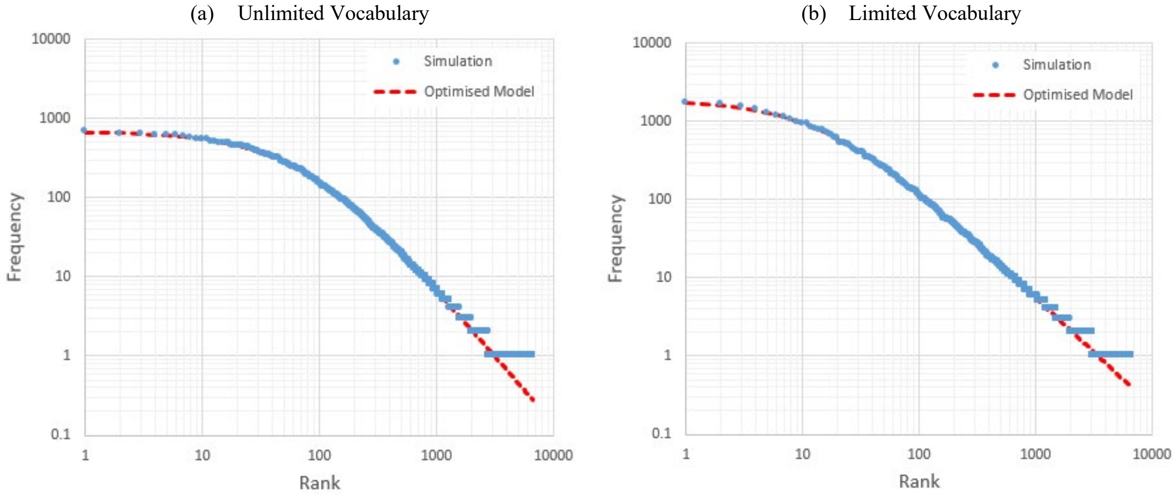
Appendix E. Model Verification by Simulation

Simulation is used to verify the mathematical model, and thus to justify the approximations used in its development. Parameters obtained from optimisation (Appendix D) are used to create a simulated version of a corpus item; each potential type is assigned a rank identifier r (not necessarily its rank in the generated document) with a selection probability given by (8). For each token, potential r-values are taken sequentially ($r = 1, 2, 3 \ldots$), and for each a uniform random number u (between 0 and 1) is generated. The first r for which $\frac{p_r}{1-\sum_{i=1}^{r-1} p_i} < u$ is selected as the type for that token. However, since selection becomes less likely as r increases, the algorithm sometimes entered an effectively infinite loop. Therefore if r reaches a pre-set value r_m without selection occurring, the inversion method was used instead: i.e. r was chosen such that $\frac{1}{1-\sum_{i=1}^{r_m} p_i} \int_{r_m}^{V} \frac{dr}{\zeta_V(\alpha,m)(r+m)^\alpha} = u$ (where again u is a uniform random variable between 0 and 1). For most our simulations we use the smallest value of r_m such that $\sum_{i=1}^{r_m} p_i > 0.99$. The resulting vocabulary profiles and frequency vs. rank and types vs. frequency distributions were compared with the corresponding model curves, and Figures E.1 to E.3 show the results, verifying that the model is approximately valid.


(a) Unlimited Vocabulary

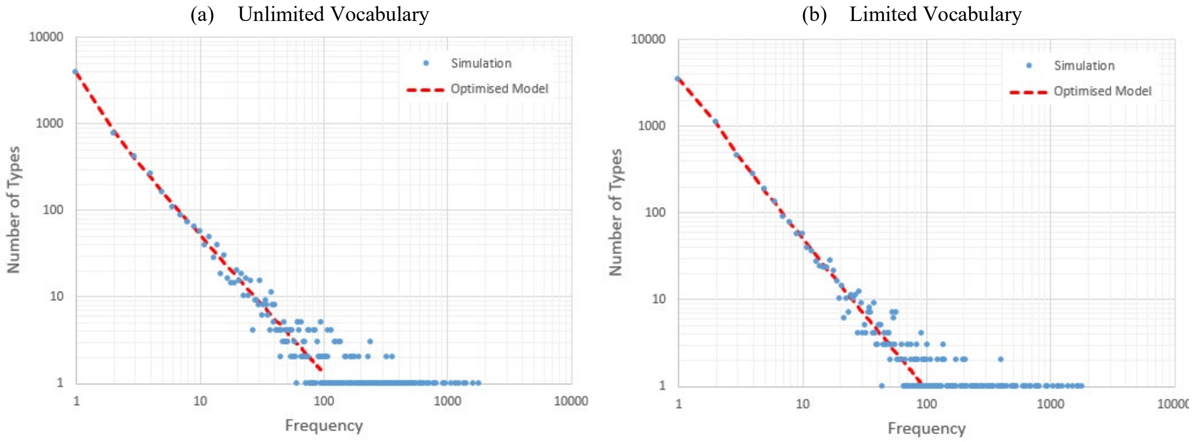
(b) Limited Vocabulary

Figure E.1: Simulated vocabulary growth curves based on the optimal model for *Three Men in a Boat*, compared with the optimal model predictions: (a) unlimited vocabulary simulation/model, (b) limited vocabulary simulation/model.


(a) Unlimited Vocabulary

(b) Limited Vocabulary

Figure E.2: Simulated frequency vs. rank distributions based on the optimal model for *Three Men in a Boat* compared with the optimal model predictions: (a) unlimited vocabulary simulation/model, (b) limited vocabulary simulation/model.


(a) Unlimited Vocabulary

(b) Limited Vocabulary

Figure E.3: Simulated types vs. frequency distributions based on the optimal model for *Three Men in a Boat* compared with the optimal model predictions: (a) unlimited vocabulary simulation/model, (b) limited vocabulary simulation/model.

References

[1] L. Wetzel, Types and Tokens", The Stanford Encyclopaedia of Philosophy (Fall 2018 Edition), E.N. Zalta (Ed.), https://plato.stanford.edu/archives/fall2018/entries/types-tokens/, 2018 (accessed 11 Nov. 2020).

[2] R.A. Fisher, A.S., Corbet, C.B. Williams, (1943), The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population, J. Animal Ecology 12 (1943) 42-58. https://doi.org/10.2307/1411.

[3] A. Orlitsky, A.T. Suresh, Y, Wu, Optimal Prediction of the Number of Unseen Species, Proc. National Academy of Sciences 113 (2016) 13283-13288. https://doi.org/10.1073/pnas.1607774113.

[4] C. Mora, D.P. Tittensor, S. Adl, G.G. Simpson, B. Worm, How Many Species are there on Earth and in the Ocean, PLoS Biology 9 (2021). https://doi.org/10.1371/journal.pbio.1001127.

[5] M.J. Costello, S. Wilson, Houlding, B, Predicting Total Global Species Richness Using Rates of Species Description and Estimates of Taxonomic Effort, Systematic Biology 61 (2012) 871-883. https://doi.org/10.1093/sysbio/syr080.

[6] E.G. Altmann, M. Gerlach, Statistical Laws in Linguistics, in: M, Degli Esposti, E. Altmann, F. Pachet (Eds), Creativity and Universality in Language: Lecture Notes in Morphogenesis. Springer, Cham, 2016, pp.7-26. https://doi.org/10.1007/978-3-319-24403-7_2.

[7] L. Lü, Z-K. Zhang, T. Zhou, T, Zipf's Law Leads to Heap's Law: Analysing Their Relation in Finite-Size Systems, PLoS ONE, 5 (2010). https://doi.org/10.1371/journal.pone.0014139.

[8] V. Davis, Types, Tokens, and Hapaxes: A Hew Heaps Law, Glottotheory International Journal of Theoretical Linguistics, 9 (2019) 113-129. https://doi.org/10.48550/arXiv.1901.00521.

[9] B. Efron, R. Thisted, Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know?, Biometrika 63 (1976) 435-447. https://doi.org/10.1093/biomet/63.3.435.

[10] G. Youmans, Measuring Lexical Style and Competence: The Type-Token Vocabulary Curve, Style, 24 (1990) 584-599.

[11] F.J. Van Droogenbroeck, Handling the Zipf Distribution in Computerized Authorship Attribution, https://www.academia.edu/24147736/Handling_the_Zipf_distribution_in_computerized_authorship_attribution, 2016 (accessed 20 July 2022).

[12] L. Quoniam, F. Balme, H. Rostaing, E. Giraud, J.M. Dou, Bibliometric Law Used for Information Retrieval, Scientometrics, 4 (1998) 83-91. https://doi.org/10.1007/BF02457969.

[13] A.M. Petersen, J. Tenenbaum, S. Havlin, H.E. Stanley, M. Perc, Languages Cool as they Expand: Allometric Scaling and the Decreasing Need for New Words, Scientific Reports 2 943 (2012). https://doi.org/10.1038/srep00943.

[14] M. Perc, Evolution of the Most Common English Words and Phrases over the Centuries, J. R. Soc. Interface 9 (2012) 3323-3328. https://doi.org/10.1098/rsif.2012.0491.

[15] M. Gerlach, F. Font-Clos, A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics, Entropy 22 (2020) 126. https://doi.org/10.3390/e22010126.

[16] G. Herdan, Type-Token Mathematics: A Textbook of Mathematical Linguistics, Mouton and Co., The Hague. 1960.

[17] L. Lü, Z-K. Zhang, T. Zhou, Deviation from Zipf's and Heaps' Laws in Human Languages with Limited Vocabulary Sizes, Scientific Reports 3 1082 (2013). https://doi.org/10.1038/srep01082.

[18] W. Dahui, L. Menghui, D. Zengru, True Reason for Zipf's Law in Language, Physica A: Statistical Mechanics and its Applications 358 (2005) 545-550. https://doi.org/10.1016/j.physa.2005.04.021.

[19] M. Brysbaert, M.A. Stevens, P. Mandera, E. Keuleers, How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age, Frontiers in Psychology, 7 (2016). https://doi.org/10.3389/fpsyg.2016.01116.

[20] A. Kornai, Zipf's Law Outside the Middle Range, Proc. 6th Meeting of Mathematics of Language, University of Central Florida, 1999, pp. 347-356.

[21] G.K. Zipf, Human Behavior and the Principle of Least Effort; an Introduction to Human Ecology, Hafner Pub. Co., New York, 1972.

[22] B. Mandelbrot, An Informational Theory of the Statistical Structure of Language, in: W. Jackson (Ed.), Communication Theory, Academic Press, Princeton, 1953, pp. 486-502.

[23] M.E.J. Newman, Power Laws, Pareto Distributions and Zipf's Law, Contemporary Physics 46 (2005) 323-351. https://doi.org/10.1080/00107510500052444.

[24] H. Bauke, Parameter Estimation for Power-Law Distributions by Maximum Likelihood Methods, European Physical Journal B 58 (2007) 167-173. https://doi.org/10.1140/epjb/e2007-00219-y.

[25] Á. Corral, I. Serra, The Brevity Law as a Scaling Law, and a Possible Origin of Zipf's Law for Word Frequencies, Entropy, 22 (2020) 224. https://doi.org/10.3390/e22020224.

[26] Á. Corral, M.G. del Muro, From Boltzmann to Zipf through Shannon and Jaynes, Entropy 22 (2020). https://doi.org/10.3390/e22020179.

[27] Z.K. Silagadze, Citations and the Zipf-Mandelbrot Law, Complex Systems, 11 (1997) 487-499. https://doi.org/10.48550/arXiv.physics/9901035.

[28] J. Nebel, S. Pezzulli, Distribution of Human Genes Observes Zipf's Law. https://eprints.kingston.ac.uk/id/eprint/44292/1/Nebel-J-C-44292-VoR.pdf, 2012 (accessed 19 Nov. 2021).

[29] G. DeMarzo, F.S. Labini, L. Pietronero, Zipf's Law for Cosmic Structures: How Large are the Greatest Structures in the Universe, Astronomy and Astrophysics 651 (2021) A114. https://doi.org/10.1051/0004-6361/202141081.

[30] P. Bak, How Nature Works: the Science of Self-Organized Criticality. Springer, New York, 1996, p.27.

[31] D. Easley, J. Kleinberg, Networks, Crowds, and Markets: Reasoning about a Highly Connected World, Cambridge University Press, 2010. Chapter 18: Power Laws and Rich Get Richer Phenomena, pp. 543-560, https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch18.pdf (accessed 18 July 2021).

[32] A. Penn, Preferential Attachment: Why The Rich Get Richer. https://www.shortform.com/blog/preferential-attachment/, 2019 (accessed 21 July 2022).

[33] F. Tria, V. Loteto, V.D.P. Servedio, Zipf's, Heaps' and Taylor's Laws are Determined by the Expansion into the Adjacent Possible, Entropy, 20 (2018) 752. https://doi.org/10.3390/e20100752.

[34] G. DeMarzo, A. Gabrielli, A. Zaccaria, L. Pietronero, Dynamical Approach to Zipf's Law, Physical Review Research 3 (2021) 013084. https://doi.org/10.1103/PhysRevResearch.3.013084.

[35] M.A. Montemurro, Beyond the Zipf-Mandelbrot Law in Quantitative Linguistics Physica A 300 (2021) 567-578, 2001. https://doi.org/10.1016/S0378-4371(01)00355-7.

[36] W. Li, Zipf's Law is Everywhere, Glottometrics, 5 (2002) 14-21.

[37] I. Moreno-Sanchez, F. Font-Clos, A. Corral, Large-Scale Analysis of Zipf's Law in English Texts, PLoS ONE 11(2016) e0147073. https://doi.org/10.1371/journal.pone.0147073.

[38] R.F. Cancho, R.V. Solé, Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited, Journal of Quantitative Linguistics, 8 (2001) 165-173. https://doi.org/10.1076/jqul.8.3.165.4101.

[39] M. Tunnicliffe, G. Hunter, The Predictive Capabilities of Mathematical Models for the Type-Token Relationship in English Language Corpora, Computer Speech & Language 70 (2021) 101227. https://doi.org/10.1016/j.csl.2021.101227.

[40] L. Boystov, A Simple Derivation of the Heaps' Law from the Generalized Zipf's Law. https://arxiv.org/pdf/1711.03066.pdf, 2017 (accessed 24 June 2022).

[41] D.C. van Leijenhorst, Th.P. van der Weide, A Formal Derivation of Heaps' Law, Information Sciences 170 (2005) 263-272. https://doi.org/10.1016/j.ins.2004.03.006.

[42] G.J.O. Jameson, The Incomplete Gamma Functions, https://www.maths.lancs.ac.uk/jameson/gammainc.pdf, 2017 (accessed 21 July 2021).

[43] I. Eliazar, The Growth Statistics of Zipfian Ensembles: Beyond Heaps' Law, Physica A: Statistical Mechanics and its Applications 390 (2011) 3189-203. https://doi.org/10.1016/j.physa.2011.05.003.

[44] F.G. Tricomi, A. Erdélyi, (1951), An Asymptotic Expansion of the Ratio of Gamma Functions, Pacific Journal of Mathematics, 1 (1951) 133-142.

[45] E.P. White, B.J. Enquist, J.L. Green, On Estimating the Exponent of Power-Law Frequency Distributions, Ecology, 89 (2008) 905-912 https://doi.org/10.1890/07-1288.1.

[46] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-Law Distributions in Empirical Data, SIAM Review 51 (2009) 661-703, 2009. https://doi.org/10.1137/070710111.