The final published version is available at https://doi.org/10.1016/j.csl.2021.101227

The Predictive Capabilities of Mathematical Models for the Type-Token Relationship in English Language Corpora

Martin Tunnicliffe[a] and **Gordon Hunter**[b]

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

Abstract

We investigate the predictive capability of mathematical models of the type-token relationship applied to the vocabulary growth profiles of selected of English language documents. We compare the existing Good-Toulmin and Heaps formulae with an alternative approach based on Bernoulli trial word selection from a fixed finite vocabulary using the Zipf and Zipf-Mandelbrot probability distributions. We make two major observations: firstly, while the Zipf-Mandelbrot model makes better predictions of vocabulary growth than the Zipf model, the optimized parameters of the latter correlate better than those of the former with statistics gleaned independently from the data. Secondly, the mean of the Zipf-Mandelbrot, Good-Toulmin and Heaps models provides a more consistent and unbiased prediction of vocabulary than any individual model alone.

Key words: types/token systems, vocabulary size, Zipf's law, Heaps' law.

1. Introduction

There have been many attempts to quantify the relationship between the number of "types" observed among a population of "tokens"; an early example was based on Corbet's 1940s survey of butterflies in the Malay peninsula (Fisher et al. 1943), types representing the species observed and tokens the individuals captured. The aim has generally been to predict how many types exist beyond those observed in a limited sample; this is called the "unseen species" problem.

Some workers have used text documents ("corpora") as models of type/token systems, with word instances as tokens and unique "vocabulary" words as types. This practice goes back at least to 1956 when Good and Toulmin (1956) used word-samples from Dickens and Macauley, and in 1976 Efron and Thisted (1976) attempted to estimate how many words Shakespeare knew. Some have even found inherent value in studying linguistic type/token systems, for example in enumerating the language development of young children (Richards, 1987). This type of study falls under what is more generally known as "statistical language modelling".

The current paper assesses the predictive abilities of several different vocabulary growth models. Each model is trained using the initial portion of each text, and then used to extrapolate the remaining unseen vocabulary. Two existing models are evaluated alongside two new models, using many different texts to achieve statistical significance. We find that some models lead to over-prediction and others to under-prediction, while an average of several models yields a fairly unbiased estimate. We further show some interesting comparisons between the optimized values of certain model parameters and the values of those same parameters independently computed.

[a] Corresponding author, e-mail: M.J.Tunnicliffe@kingston.ac.uk.
[b] E-mail: G.Hunter@kingston.ac.uk

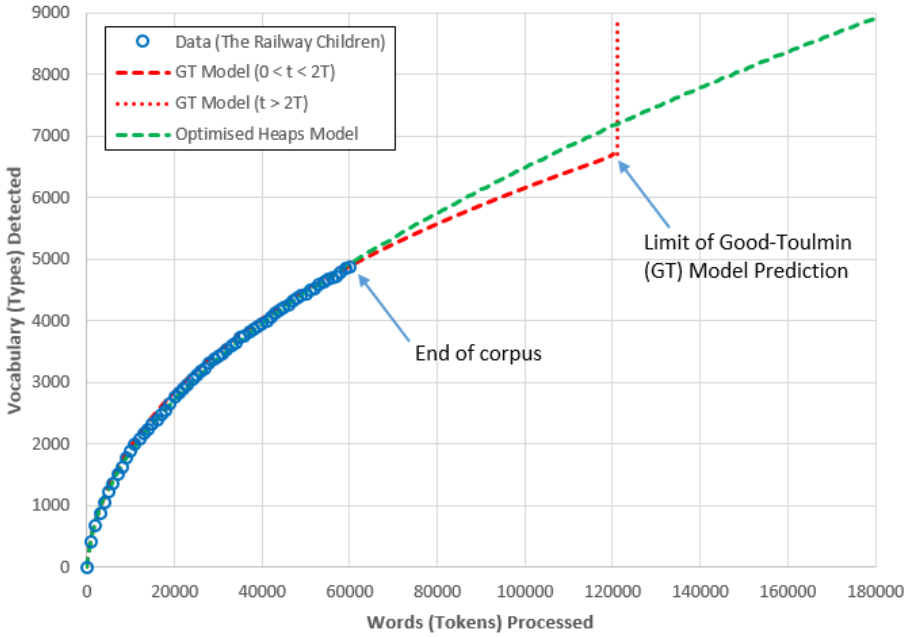

Figure 1: Good-Toulmin (GT) and Heaps models optimized to the novel *The Railway Children* showing the extrapolated vocabulary growth had the story continued beyond its current ending.

2. Problem Definition

We define t as the number of tokens observed up to an arbitrary point during processing, and $v(t)$ as the corresponding number of types (the "vocabulary to date"). We further define T as *total* number of tokens at the end of processing, such that the total vocabulary $V = v(T)$. Figure 1 shows the $v(t)$ profile for a typical novel, processed using the methodology described in Section 4. Growth is initially rapid, since here nearly all word-types are being encountered for the first time. However, the increasing frequency of "already seen" words gradually slows the vocabulary growth. While occasional departures from this rule do exist (see later), this represents by far the most frequently observed vocabulary growth behaviour, and is reflected in the models described in Section 3.1.

We use a large number of different texts in order to assess a range of vocabulary growth models, whose predictive capabilities are tested by: (i) optimizing them to fit the first half of each corpus, (ii) predicting subsequent vocabulary growth across the second half and (iii) comparing the results with the entire corpus vocabulary. We also make comparisons between the optimized model parameters, and independent measurements of those same parameters.

3. Related Work

3.1 Vocabulary Growth Models

One of the earliest mathematical studies of type/token growth concerned A.S. Corbet's survey of Malaysian butterflies in the early 1940s (Fisher et al., 1943). Having counted the number of species g_f for which he had captured exactly f individuals, Corbet wished to estimate how many additional species would be found in a second sample of equal size. His co-author Fisher proposed a solution which was subsequently generalized by Good and Toulmin (1956) for unequal sample sizes; using our own notation, we can write their formula $v(t) = \sum_{f=1}^{\infty} g_f \left[1 - (1 - t/T)^f\right]$, a concise derivation of which can be found in Efron and Thisted (1976). (Recently, an equivalent expression was independently

derived by Font-Clos and Corral (2015), albeit with a Zipfian distribution in the place of g_f.) We refer to this as the "Good-Toulmin" or GT model.

GT models past vocabulary growth ($0 \leq t \leq T$) and extrapolates future behaviour across the range $T \leq t \leq 2T$. Unfortunately, for $t > 2T$ the summation fails to converge and the formula becomes useless (see Figure 1). There have been many attempts to tweak and reformulate the problem in order to overcome this problem (e.g. Orlitsky et al., 2016) and potentially predict the *total* number of types ($v(\infty)$). One example was by Efron and Thisted (1976) who deduced that Shakespeare knew around 35,000 words that he never used in any of his plays and poems - a claim which will obviously never be tested!

A considerably simpler formula was discovered empirically by Heaps and independently by Herdan (Egge, 2007): $v(t) = At^{\lambda}$ where A and λ are constants obtained by optimizing the model to an observed profile ($0 \leq t \leq T$). Such a curve has been added to Figure 1; while it lacks the range limitation of GT, it predicts (for this corpus) a significantly faster and ultimately unbounded vocabulary growth. (Font-Clos and Corral (2015) however report a log-log convexity in the type/token curve, suggesting that growth is ultimately slower than this model would require.) We refer to this model as "Heaps' Law" or simply "Heaps".

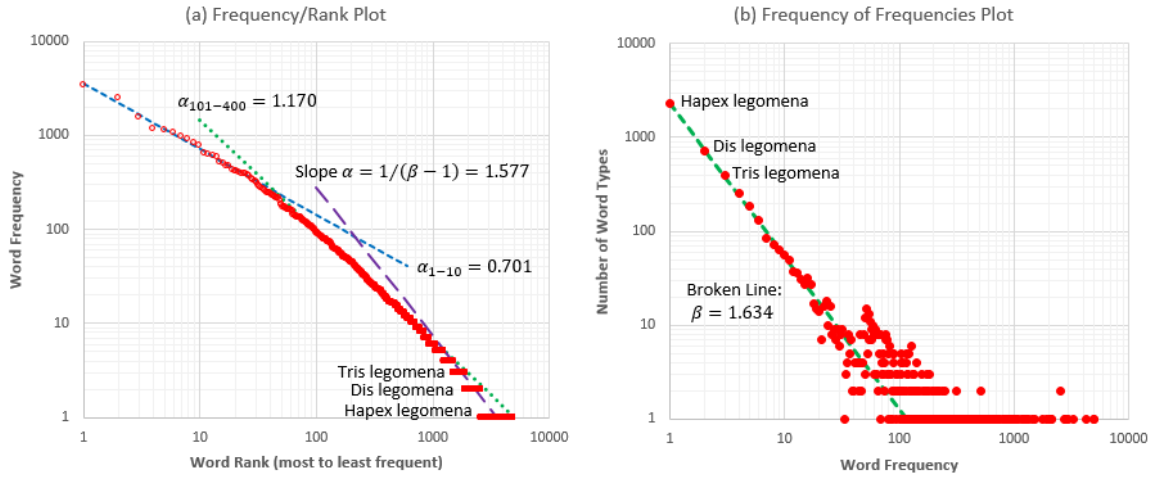

Figure 2: (a) Frequency vs. rank and (b) "frequency of frequencies" distributions for the novel *The Railway Children* (see Section 3) showing hapex, dis and tris legomena (words appearing only once, twice and three times).

3.2 The Zipf and Zipf-Mandelbrot Laws

Heaps' law is often associated with another empirical type-token relationship: the generalised Zipf's law (Zipf, 1949) which states that the frequency of a type is related to its "rank" r: $r = 1$ being the most frequent type, $r = 2$ the next most frequent etc. The frequency f_r (number of tokens) of the type is related to its rank by $f_r \propto 1/r^{\alpha}$ where α is usually called the "Zipf index".

Figure 2(a) shows typical frequency vs. rank data (again based on *The Railway Children*) plotted on a log-log scale. The slope (Zipf index) is not quite constant, though the graph can be divided into quasi-linear "high frequency" and "low frequency" domains, the latter having a higher slope than the former. Mandelbrot (1953) proposed a further generalization of Zipf's law to embrace this effect, commonly referred to as the "Zipf-Mandelbrot" law $f_r \propto 1/(r+m)^{\alpha}$ where m is an additional constant. At the low-frequency tail of the graph, the *hapex* (and respectively *dis*, *tris* etc.) *legomena* - words appearing

only once, twice, three times etc. - appear arbitrarily ranked; these can be arranged as a "frequency of frequencies" distribution (Figure 2(b)) which displays a Zipf-like relationship of its own: $g_f \propto 1/f^{\beta}$ where g_f is the number of types of which exactly f tokens exist. This is sometimes referred to as Zipf's "second law". It has been argued (Lü et al., 2010) that two Zipf indices are related by $\alpha = 1/(\beta - 1)$ and the slope predicted by this formula is superimposed on Figure 2(a), showing a plausible agreement in the low-frequency tail.

Several workers have shown analytically that Heaps' law is a consequence of Zipf's, with a simple inverse relationship between the two indices ($\lambda = 1/\alpha$). Boystov (2017) (originally published in Russian in 2003), assuming Bernoulli trial word-selection from an infinite vocabulary with a Zipfian probability distribution, showed that $v(t) = O\left(t^{1/\alpha}\right)$, a result obtained independently by van Leijenhorst and van der Weide (2005) from the the Zif-Mandelbrot law. Starting from from Zipf's second law, Lü et al. (2010) derived a an implicit equation for the vocabulary profile $t = v\,(1 - v^{\alpha-1})/(1 - \alpha)$ which is asymptotically equivalent to Heaps' law with $\lambda \approx \max(1/\alpha\,, 1)$. This makes qualitative sense considering that small α implies that nearly all words are unique and vocabulary grows approximately linearly with t.

3.3 Applications

The ontological distinction between general *types* of things and their specific concrete *tokens* has long been a concern of philosophers (Wetzel, 2018) and is important in both the natural and artificial domains. An improved model of the type/token profile could benefit many different fields; as well as the aforementioned studies of children's vocabulary (Richards, 1987), type-token graphs have been used to compare the lexical richness of documents (Hussain, 2015; Youmans, 1990) and as "style markers" to indicate authorship (Rudman, 2000). In the context of applications to Automatic Speech Recognition (ASR) Systems, Hunter and Huckvale (2006) and Hunter (2004) investigated how the "coverage", i.e. the fraction of word token instances in the large British National Corpus (BNC) accounted for by a finite vocabulary, depended on the size of the vocabulary considered, and how the rankings of the most common words compared between the "text" and "dialogue" portions of the BNC. Moore (2001, 2003) and Deng and Horvitz (quoted in Huang et al., 2014) investigated the dependence of "Word Error Rate" for ASR systems on the size of the training corpus used. Another application is the relative impact of musical compositions, compositions themselves being the types and performances the tokens (Dodd and Letts, 2017). Biometric studies make extensive use of type/token theory; observed patterns indicate that there are around 8.7 million eukaryotic species on Earth, but that the vast majority of these still await identification and classification (Mora et al., 2011). Costello et al. (2012) note a reduction in the rate of species discovery since the 1920s, suggestive of the gradual sparsening of undiscovered types seen in Figure 1. However, until a group inventory is nearly complete, estimating its eventual size is extremely difficult (Bebber et al., 2007).

It should be noted that in the real world uncertainties exist concerning what constitutes a "type". For example, the same word may have multiple different interpretations: "router" for example could refer to a workshop tool as well as to a piece of telecommunications equipment: if these are to be classified as different types, context will need to be considered. Also, specific words for are classified into more general groups such as nouns, prepositions, verbs etc. which could also be considered "types". Similarly biological types could be represented not only by species but also genera, orders and phyla (Mora et al, 2011), and there are nearly always sub-types within more general classifications. (For example the domestic cat *felis catus* includes Siamese, Burmese, Persian, Manx as well as mixtures of these breeds.) However, this difficulty lies outside the scope of the current paper.

Table 1: List of sources used in this study, with author in parenthesis. (Year indicates year of first publication. The parameter ρ quantifies to vocabulary size relative to the document length – see text for a full explanation.)

#	Description	Year	T (Tokens)	V (Types)	ρ
1	Three Men in a Boat (Jerome K. Jerome)	1889	68809	6593	0.04996
2	Sense and Sensibility (Jane Austen)	1811	119958	6418	-0.12016
3	The Picture of Dorian Gray (Oscar Wilde)	1890	79825	6896	0.027144
4	The Catcher in the Rye (J.D. Salenger)	1945	74270	4202	-0.16744
5	The Song of Hiawatha (Henry Wadsworth Longfellow)	1855	32387	3858	0.032047
6	The Hobbit (J.R.R. Tolkien)	1937	96837	6067	-0.08355
7	Treasure Island (Robert Louis Stephenson)	1881	68705	6021	0.010977
8	Mansfield Park (Jane Austen)	1814	160555	7926	-0.11159
9	Pride and Prejudice (Jane Austen)	1813	122197	6379	-0.12807
10	Friends Season 1 Episodes 1-8 (various authors)	1994	72342	4715	-0.10991
11	The Great Gatsby (F. Scott Fitzgerald)	1925	49037	5945	0.101592
12	Tom Sawyer (Mark Twain)	1876	71379	7427	0.091238
13	Huckleberry Finn (Mark Twain)	1884	112341	6347	-0.10629
14	The Girl in the Golden Atom (Ray Cummings)	1920	97073	6211	-0.07406
15	The War of the Worlds (H.G. Wells)	1897	60882	6741	0.094491
16	Nineteen Eighty Four (George Orwell)	1949	105382	9209	0.073585
17	Just William (Richmal Crompton)	1922	47955	5433	0.06884
18	Of Mice and Men (John Steinbeck)	1937	29339	2952	-0.05603
19	The Pilgrim's Regress (C.S. Lewis)	1933	67718	6513	0.049214
20	Brideshead Revisited (Evelyn Waugh)	1945	117636	10934	0.116795
21	Gulliver's Travels (Jonathan Swift)	1726	104910	8094	0.018815
22	The Wonderful Wizard of Oz (E. Frank Baum)	1900	39578	2951	-0.1415
23	Wuthering Heights (Emily Brontë)	1847	119404	9759	0.063169
24	The Railway Children (Edith Nesbit)	1906	60174	4887	-0.04186
25	The Man Who Was Thursday (G.K. Chesterton)	1908	58930	6316	0.075498
26	Hard Times (Charles Dickens)	1854	105692	8754	0.050742
27	The Mayor of Casterbridge (Thomas Hardy)	1886	117840	10278	0.089431
28	The Princess and Curdie (George Macdonald)	1883	56961	5024	-0.01421
29	Alice's Adventures in Wonderland (Lewis Carroll)	1865	27317	2772	-0.06299
30	Missee Lee (Arthur Ransome)	1941	97321	5019	-0.16733
31	The Wind in the Willows (Kenneth Graham)	1908	59804	6569	0.088358
32	The Sign of Four (Arthur Conan Doyle)	1890	43634	5489	0.10021
33	Howard's End (E.M. Forster)	1910	110257	9259	0.063046
34	Lord Jim (Joseph Conrad)	1900	133730	10492	0.062323
35	Uncle Tom's Cabin (Harriet Beecher Stow)	1852	188008	11502	0.005132
36	The Jungle Book (Rudyard Kipling)	1894	52300	4579	-0.03015
37	Tarzan of the Apes (Edgar Rice Burrows)	1912	86097	7429	0.037916
38	Little Lord Fauntleroy (Frances Hodgson Burnett)	1885	59004	4661	-0.05682
39	Frankenstein (Mary Shelley)	1818	75314	6979	0.048921
40	Robinson Crusoe (Daniel Defoe)	1719	121741	6020	-0.15216
41	Tom Brown's Schooldays (Thomas Hughes)	1857	107107	8946	0.056373
42	The Life and Opinions of Tristram Shandy (Laurence Sterne)	1759	189266	14425	0.101574
43	The Rainbow (D.H. Lawrence)	1915	187678	11240	-0.00437
44	Little Women (Louisa May Alcott)	1868	187068	11423	0.003568
45	The Pilgrim's Progress (John Bunyan)	1678	57446	4391	-0.07511
46	Pollyanna (Eleanor H. Porter)	1913	57062	4331	-0.07917
47	Ulysses (James Joyce)	1922	269868	28998	0.303696
48	Brave New World (Aldous Huxley)	1931	66618	8432	0.166032
49	Moby Dick (Herman Melville)	1851	216217	17503	0.147622
50	The Book of Mormon (Joseph Smith tr.)	1830	275887	5616	-0.41553

4. Materials and Methods

4.1 Source Data

The fifty texts listed in Table 1 were selected with a view to creating a representative sample of typical English fictional or literary documents. Most of these are novels (adult, juvenile, classic, pulp, U.S. and U.K.) with a smattering of short stories, poetry, popular drama, religious allegory and scripture. The earliest is John Bunyan's *The Pilgrim's Progress* (1679) and the most recent the scripts for first series of the US TV sitcom *Friends* (1994). All are original English compositions, the one arguable exception being *The Book of Mormon*, claimed by Latter-day Saints to be a translation of an ancient historical record. This was included on account of certain interesting properties which will be discussed later.

Each document was processed electronically using a purpose-written C++ program to discover its size T (total number of tokens/words) and observed vocabulary V (types/ unique words). The program also recorded the vocabulary growth profile $v(t)$ for $t = 1000, 2000, 3000, \ldots$ processed words (e.g. Figure 1) and the word frequency/rank distributions (e.g. Figure 2). A word "type" was defined as any unique sequence of letters, regardless of any common stem: for example "boy", "boys", "boy's", "boyish" and "boyhood" all counted as distinct types. Numerals were ignored, as was all punctuation with the exception of the apostrophe (which was treated as a letter).

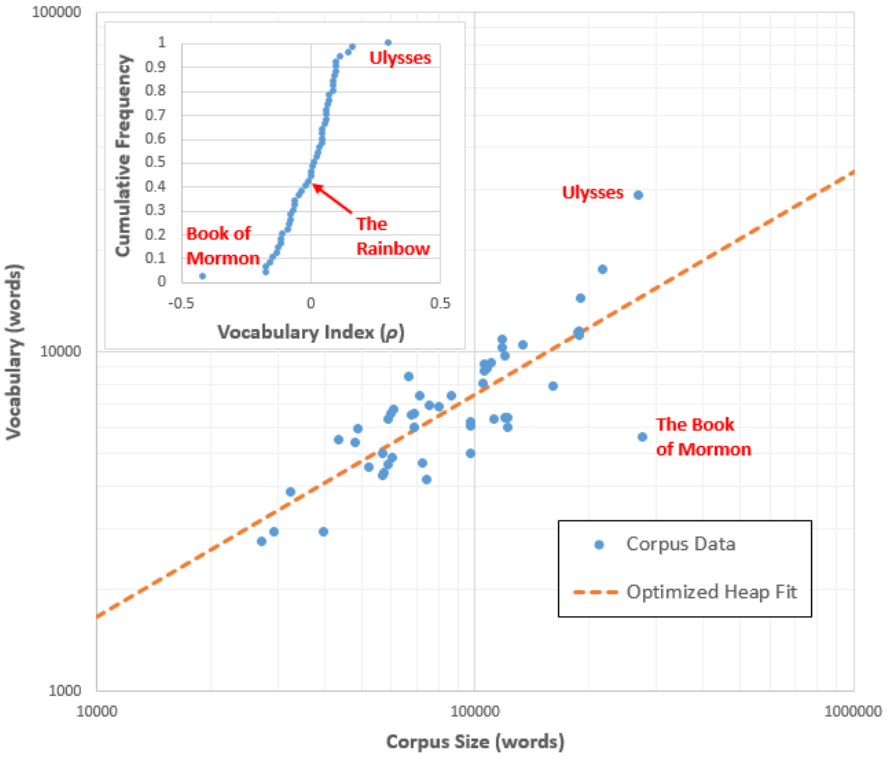

Figure 3: Log-log relationship between corpus size and vocabulary, with outliers highlighted. Inset graph shows the cumulative distribution of the vocabulary index.

V is clearly not a useful measure of the linguistic diversity of a corpus since it depends on the document length T. The type/token ratio $TTR = V/T$ has sometimes been used (Richards, 1987) but is unsatisfactory since it also depends upon T. However, plotting V vs. T on a log-log graph (Figure 3) reveals a trend similar to that of Heaps' law, so we may define $\bar{V}(T) = AT^{\lambda}$ where optimal values of A and λ may be obtained using linear regression in the log-log domain. The resulting residuals may then

be seen to represent the richness/paucity of the individual texts' vocabularies relative to the group, so we define the "vocabulary index"

$$\rho(V,T) = \log_{10}\frac{V}{\bar{V}(T)}, \tag{1}$$

recognizing of course that this is only a relative measure within this particular dataset. The inset on Figure 3 shows that the distribution of ρ is approximately uniform with two notable outliers: (i) James Joyce's *Ulysses* has by far the largest index (0.304) due to its many "nonce" words (coined by the author for a single use, e.g. *poppysmic*, *ringroundabout*, *yogibogeybox*…) and (ii) *The Book of Mormon* which, though it has by no means the smallest vocabulary, has by far the lowest index (−0.42) due to its frequent repetitions ("And it came to pass…" etc.). D.H. Lawrence's *The Rainbow* lies close to the median, with $\rho \approx 0$.

The "high" and "low" frequency Zipf indices were measured for all 50 corpora, by applying linear regression (in the log-log domain) to the rank-ranges 1-10 and 101-400 respectively. Although the frequency/rank plot of Figure 2 is certainly typical, not all distributions were so "well behaved" (see Figure 4). Nevertheless, as Figure 5(a) shows, the high frequency indices were nearly always significantly smaller than their low frequency counterparts (the one exception being *The Song of Hiawatha*). Interestingly, the two α-values have a weak but very significant negative correlation which meets the criterion for 99.5% confidence (p-value < 0.005), suggesting that writers working with fewer infrequent words tend to select their most common words more evenly.

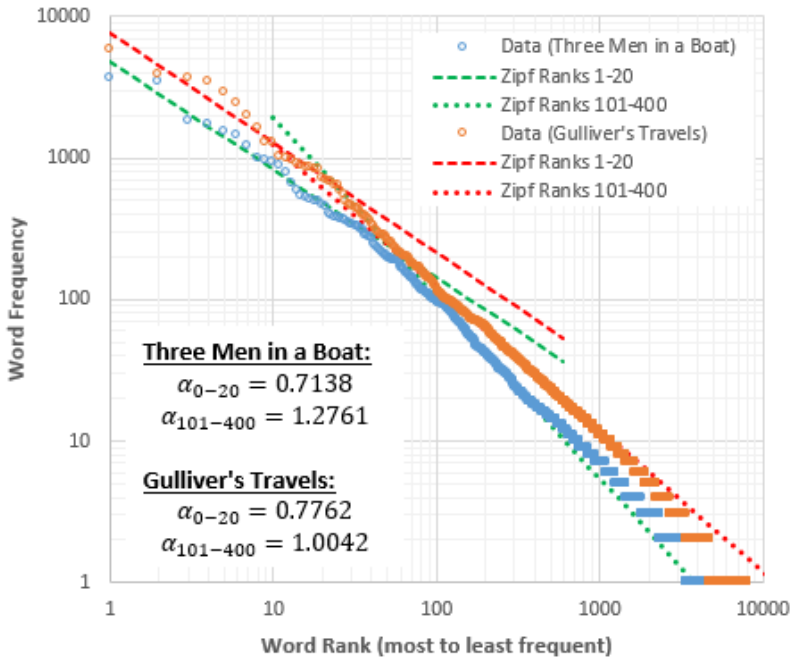

Figure 4: Zipf word frequency plots for the novels *Three Men in a Boat* and *Gulliver's Travels*. Although Zipf's law does not always apply precisely, we nevertheless use the α-values obtained by linear regression in the log-log domain over the rank-ranges 1-20 and 101-400 to characterize the distributions.

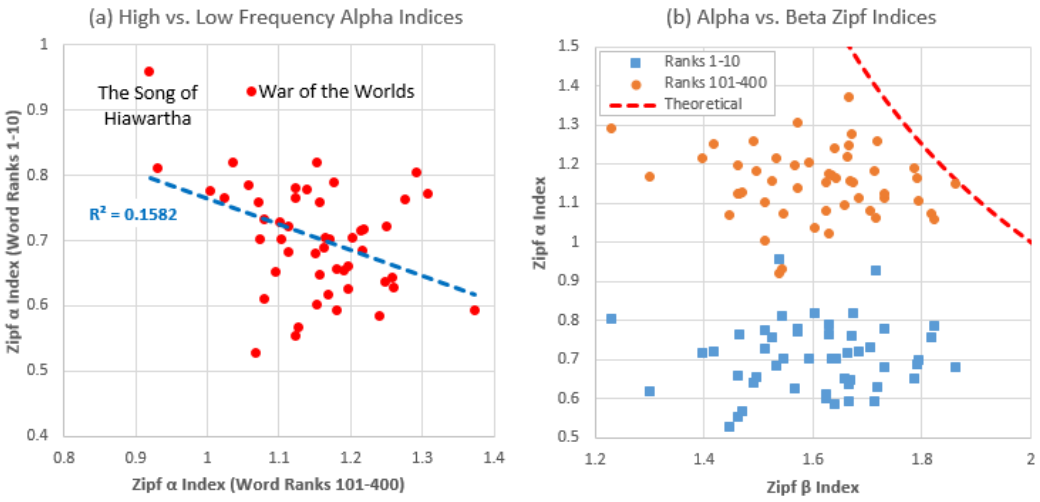

Figure 5: (a) Low and high-frequency Zipf indices show a small but significant negative correlation. (b) Theory predicts larger α-values than were observed, given the measured β-values.

Figure 5(b) compares the observed α vs. β plots, which show no significant correlation or clear agreement with the $\alpha = 1/(\beta - 1)$ "theoretical" curve (Lü et al. 2010). One possible explanation is that the "true" α within the low frequency (*hapex legomena,* etc.) tail is much larger than that measured directly for word ranks 101-400 and that its true value can only be computed indirectly (as in Figure 2).

4.2 Model Optimization

With the exception of the GT model (which is parameterized by independently observed frequencies-of-frequencies g_f) we optimized our models on observed vocabulary profiles and considered (i) how well the model fit the data, (ii) how closely the optimized parameters agreed with independent observations and (iii) how well the optimized model predicted future vocabulary development. For every Δt processed tokens, the number of types $\hat{v}(i\Delta t)$ was recorded, for $i = 1, 2, 3 \ldots \lfloor T/\Delta t \rfloor$ (T being the total tokens processed and $\lfloor x \rfloor$ the largest integer not exceeding x). We began with an initial model $\Pi(p_1, p_2 \ldots)$ where $p_1, p_2 \ldots$ are adjustable parameters, which yielded a vocabulary profile $v_\Pi(t)$. We then applied the following procedure:

1. Calculate the total squared error $\varepsilon(\Pi) = \sum_{i=1}^{\lfloor T/\Delta t \rfloor} [v_\Pi(i\Delta t) - \hat{v}(i\Delta t)]^2$.
2. Propose a "mutated" hypothesis $\Pi'(p_1[1 + \Delta x], p_2[1 + \Delta y] \ldots)$ where Δx, Δy etc. are dimensionless numbers chosen according to some scheme (see below).
3. Calculate the mutated total squared error $\varepsilon(\Pi') = \sum_{i=1}^{\lfloor T/\Delta t \rfloor} [v_{\Pi'}(i\Delta t) - \hat{v}(i\Delta t)]^2$.
4. If $\varepsilon(\Pi') < \varepsilon(\Pi)$ then accept the mutation as "beneficial" and make $\Pi = \Pi'$. Otherwise keep the existing model and return to step 2.

This procedure produced progressively better and better models. (Throughout this work, Δt was set arbitrarily to 1000 tokens.)

The following policy was used to select Δx, Δy,…, etc. for each step:

1. Choose a dimensionless "step size" S (see below).
2. Compute the partial derivatives $\frac{\partial \varepsilon}{\partial x}$, $\frac{\partial \varepsilon}{\partial y}$… etc. (either analytically or using a finite difference approach).

3. Compute a first mutation $\Delta x_1 = -\frac{\partial \mathcal{E}}{\partial x}\frac{S}{P}$, $\Delta y_1 = -\frac{\partial \mathcal{E}}{\partial y}\frac{S}{P}\ldots$, where $P = \sqrt{\left(\frac{\partial \mathcal{E}}{\partial x}\right)^2 + \left(\frac{\partial \mathcal{E}}{\partial y}\right)^2 + \cdots}$.
4. Create a "semi-mutated" model $\Pi''(p_1[1+\Delta x_1], p_2[1+\Delta y_1] \ldots)$ and compute its partial derivatives.
5. Repeat step 3 to obtain a second mutation Δx_2 and Δy_2.
6. Add and normalize the two mutations, i.e. $\Delta x = (\Delta x_1 + \Delta x_2)/Q$, $\Delta y = (\Delta y_1 + \Delta y_2)/Q$,…, etc. where $Q = \sqrt{(\Delta x_1 + \Delta x_2)^2 + (\Delta y_1 + \Delta y_2)^2 + \cdots}$.
7. Construct the "fully-mutated" model $\Pi'(p_1[1+\Delta x], p_2[1+\Delta y] \ldots)$.

A policy was also needed to select the step-size S: we began with $S = 0.1$ and adjusted it after each iteration as follows:

1. If the previous/most recent mutation was beneficial, S was increased by 10%,
2. If the previous/most recent mutation was not beneficial, S was decreased by 50%.

This is clearly akin to the "additive increase, multiplicative decrease" (AIMD) algorithm used for congestion window control in TCP, and has similar benefits (Chiu and Jain, 1989). The idea was to drive the model at a cautiously accelerating rate towards its optimum, while at the same time radically slowing its evolution when overshoot was detected.

Finally a policy was needed to detect when the optimum has been found and thus to terminate the procedure at this point. This was selected quite arbitrarily: if S decreased below 10^{-10} or did not exceed 10^{-5} for 100 consecutive beneficial mutations, then the model was considered optimal.

The authors do not claim that that this is necessarily the best optimization procedure possible, only that it produces sensible results consistent with those of other less efficient schemes which we tried. For the different models that were tested, the final optimisations were found not to depend on the initial conditions, suggesting that the procedure does indeed find genuine global optima.

5. Good-Toulmin and Heaps Models

These two models have already been introduced in Section 2.1. The many parameters of the GT model $\Pi_{gt}(T, g_1, g_2, \ldots, g_{f_{max}})$ are measured directly from the word frequency distribution and are not subject to iterative optimization; the vocabulary profile is given by

$$v_{\Pi_{gt}}(t) = \sum_{f=1}^{f_{max}} g_f \left[1 - \left(1 - \frac{t}{T}\right)^f\right];\ 0 \le t \le 2T \tag{2}$$

.

where f_{max} is the largest observed word frequency. For simplicity we limit ourselves to the original formulation, and thus to the range $0 \le t \le 2T$; though somewhat restrictive, it provides reasonable scope to compare GT predictions with those of other models. The Heaps model $\Pi_h(A, \lambda)$ has only two parameters which must be chosen such that

$$v_{\Pi_h}(t) = At^{\lambda} \tag{3}$$

optimally fits the measured vocabulary profile $\hat{v}(t)$. Though this may be achieved by linear regression in the log-log domain, the procedure outlined in Section 3.2 was found to produce a smaller mean square

error and was therefore adopted. It was discovered that the final solution was not significantly affected by the initial assumptions: for the reported results however, the initial model was always $\Pi_h(0.5,5)$.

5.1 Optimizing Model using Entire Corpora

Initially each model was applied to each entire corpus to determine the quality of the overall fit. This was quantified by the RMS error

$$\epsilon_\Pi = \sqrt{\sum_{i=1}^{\lfloor T/\Delta t \rfloor} \left(v_\Pi(i\Delta t) - \hat{v}(i\Delta t)\right)^2 / \lfloor T/\Delta t \rfloor} \tag{4}$$

where T is the corpus size in words and the counting interval Δt is set to 1,000 words. We have already shown both these models applied to the entire text of *The Railway Children* (Figure 1) indicating a close qualitative agreement with the data; it is only in their future predictions that the two models significantly disagree.

However, even within the observed data range the two models do not always agree so well: profiles for *Ulysses* and *The Book of Mormon*, the two major outliers in Figure 3, are shown in Figure 6. Both are atypical of the dataset (of which Figure 1 is much more representative) with significant variations in the rate of vocabulary growth; however, both show a qualitatively superior fit for the Heaps model over the GT model, a trend which is borne out in the RMS error data in Figure 7(a). Here on average the Heaps model fits the data with about half the RMS error of the GT model, though with the one major outlier *Ulysses* removed the errors of the two models are not significantly correlated ($p = 0.1065$). Figure 7(b) shows the cumulative distribution of the difference between the two models' errors, showing that in around 80% of cases Heaps fits the data better.

The optimized values for the Heaps coefficient λ were compared with previously measured parameters for the same corpora. Figure 8 shows λ plotted against the Zipf index α calculated from the high and low-frequency words and also computed indirectly from the β index; of these, only the latter shows strong correlation, meeting the 99% confidence ($p < 0.01$) criterion for a relationship of inverse proportionality, and a plausible qualitative agreement with $\lambda = 1/\alpha$.

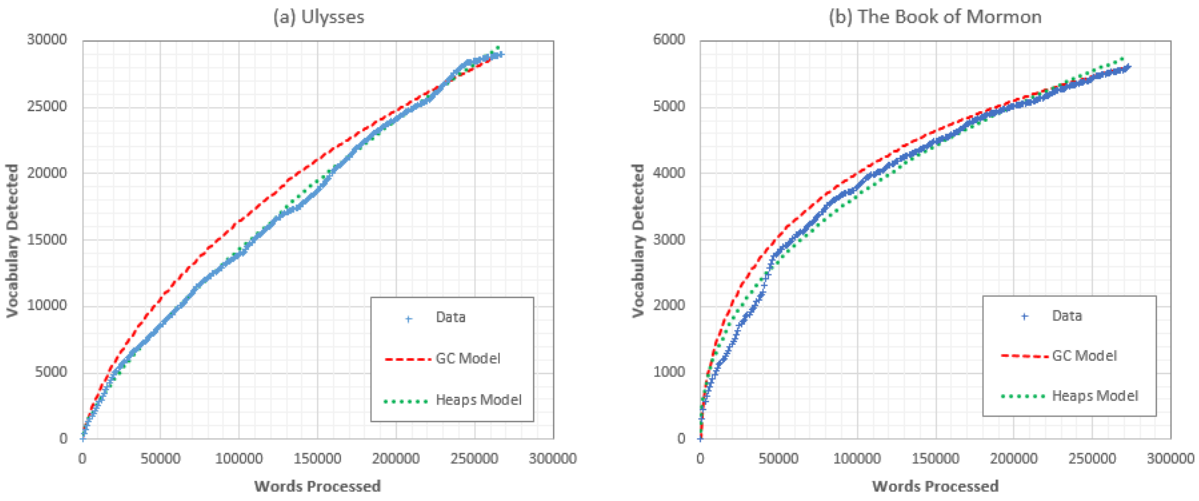

Figure 6: Measured vocabulary growth data for corpora (a) *Ulysses* and (b) *The Book of Mormon*, compared with those of the corresponding GT and optimised Heaps models.

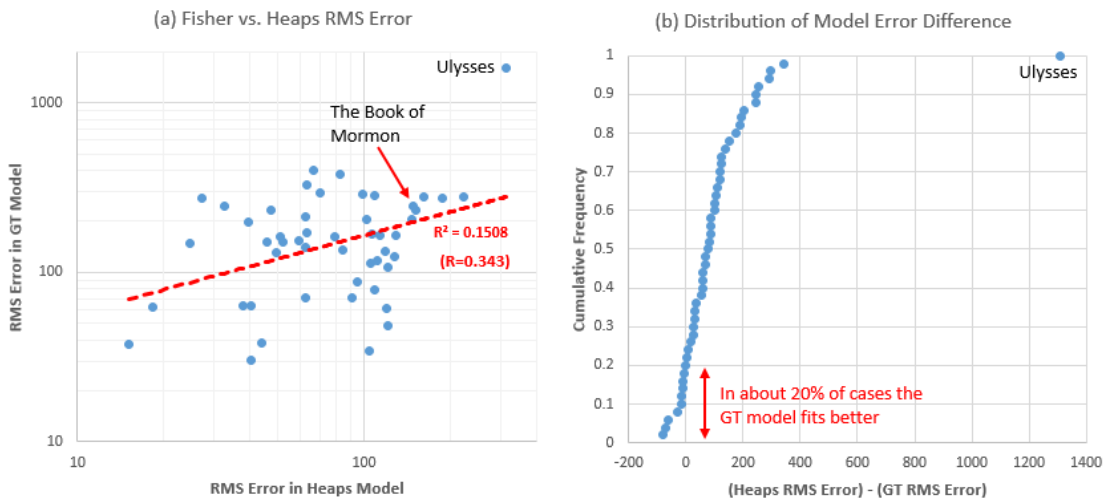

Figure 7: Comparison of the RMS errors for GT and Heaps models. (a) A weak positive correlation exists, but this is largely due to one major outlier (*Ulysses*) which produces an error significantly greater than average in both models. Removal of this point reduces the Pearson coefficient from 0.343 to 0.231, narrowly missing the criterion for 90% confidence (0.235 for $p = 0.1$). (b) The cumulative distribution of the error difference shows that in about 80% of cases, Heaps produces the better fit.

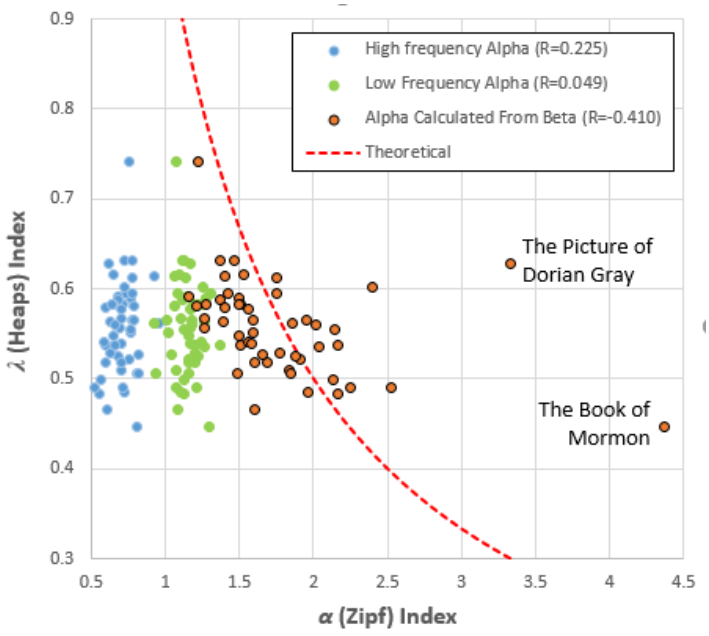

Figure 8: Optimized Heaps index plotted against Zipf α indices measured for high frequency words (ranks 1-10), low frequency words (ranks 101-400), and calculated indirectly using $\alpha = 1/(\beta - 1)$. Though the first two show no significant correlation, the latter has a strong negative correlation, meeting the 99.5% confidence criterion ($p < 0.0031$). It also suggests a plausible qualitative agreement with the theoretical $\lambda = 1/\alpha$.

5.2 Extrapolation from Partial Corpora

Figure 1 shows how a model, once optimized (or "trained") using existing data, can be used extrapolate future vocabulary growth. Though it is impossible to test how the vocabulary of (for example) *The Railway Children* would have developed had Edith Nesbit continued the story, it is nevertheless possible to assess the predictive capability of a model as follows:

1. Optimise the model using the first portion of a corpus,
2. Extrapolate this model to predict the remainder of the vocabulary profile,
3. Compare this prediction with the actual profile of the remainder of the corpus.

Since the basic GT model cannot predict beyond twice the length of the original document (see Section 3), the *first half* of each corpus was used for training and the second half for testing predictions. Figure 9 shows two examples: in both cases (as in Figure 1) the Heaps model predicts a faster vocabulary growth than the GT model, the real data lying between the two predicted profiles.

Figure 10 shows summary results for all 50 corpora, showing that the Heaps and GT models do generally over- and under-predict respectively, and that the simple arithmetic mean of the two provides a plausible unbiased estimator. The inset on the bottom right shows the mean percentage error of the two models and their mean, with error bars indicating the 95% confidence interval (i.e. $\pm 1.98\sigma/\sqrt{50}$) for the mean. The error ranges of the individual models are clearly above and below the zero line, but the error range of their average embraces the zero line, suggesting that this could indeed be an unbiased estimator (though with the suggestion of a possible small bias towards overestimation).

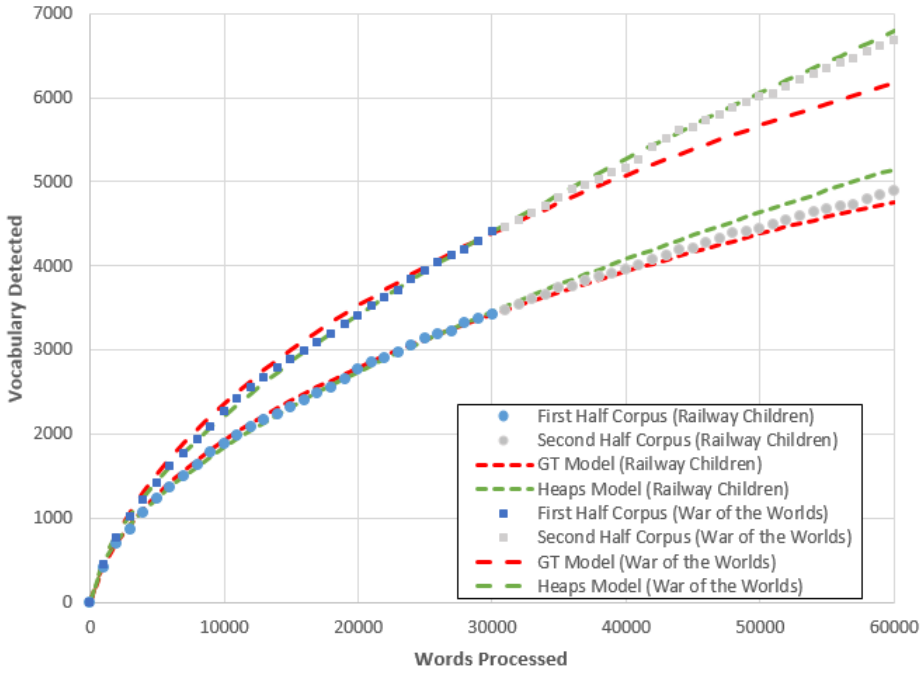

Figure 9: Vocabulary forecasts produced by GT and Heaps models, optimised using the first half of each profile and used to predict the second half; the prediction was then compared with the actual second half. The novels *The Railway Children* and *The War of the Worlds* were chosen on account of their similar lengths ($\approx$ 60,000 words) and different vocabulary indices (-0.042 and 0.094 respectively).

6. Alternative Models

Having tested and compared two existing models for vocabulary growth, we proceed to develop our own. We start by stating and attempting to justify our assumptions.

6.1 Modelling Assumptions

Our first assumption is that the available vocabulary behind each corpus is finite and constant, a statement which cannot be proven empirically and which which contradicts the views of some other

workers (Altmann et al., 2009). It is also incompatible with the Heaps model which requires a non-saturating increase in vocabulary with corpus expansion; however, since the latter yields significant over-predictions (see Figure 10) this may not in itself be a reason to discount our assumption.

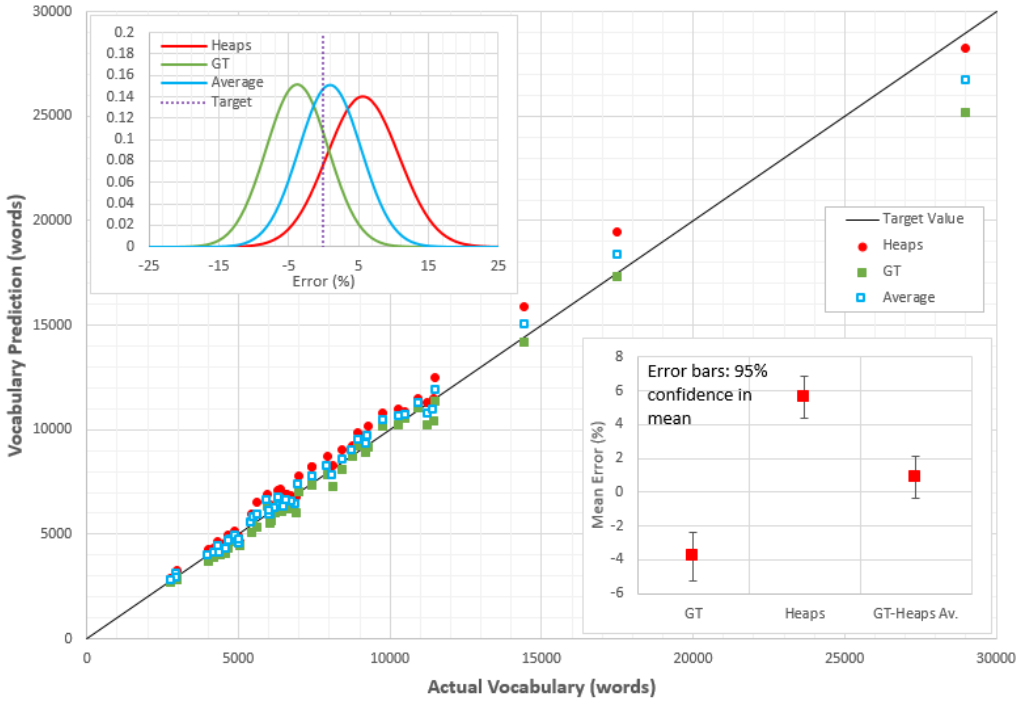

Figure 10: Total corpus vocabularies predicted from the first 50% of each corpus using the Heaps and GT models and the average of the two. Top left inset shows optimized Gaussians based on the mean and standard deviation errors of the two models. Bottom right inset shows the corresponding 95% error bars indicating that the average of the two models is a plausible unbiased estimator.

Proceeding from this assumption we adopt the approach of Boystov (2017) and van Leijenhorst and van der Weide (2005) by representing corpora as random samples from Bernoulli processes, i.e. words are selected independently, each word-type w having a characteristic probability p_w. While inter-word correlation and the "word clustering" effect (Font-Clos and Corral, 2015) prevents this from being true over short passages, it may nevertheless provide an acceptable approximation over larger scales. A Bernoulli process implies a geometric distribution for the separation k_w between any two instances of w, with expectation $E[k_w] = 1/p_w$ and coefficient of variation (i.e. the mean divided by the standard deviation) $c_w = \sqrt{1 - p_w}$. If p_w is small ($\lesssim 0.01$) this is practically indistinguishable from 1, so the difference between the observed correlation coefficient and unity provides an approximate measure of how well a particular word is modelled by this Bernoulli process.

Figure 11 shows the coefficients of variation for the 400 most frequent words in three novels, plotted against the mean word separation and as normalized frequency distributions. Although the coefficients are widely scattered, the distributions are almost identical with c_k mostly > 1. However, the modes are all close to 1, suggesting a predilection towards Bernoulli behaviour. Furthermore, it has been noted by Font-Clos and Corral (2015) that the first appearances of words (which alone affect vocabulary growth/size) follow very nearly an exponential distribution which is the limiting case of Bernoulli as $c \longrightarrow 1$. We therefore proceed with Bernoulli as a working assumption. Interestingly, a weak negative correlation exists between c_k and mean word separation, suggesting that less frequent words are spaced more regularly; however, only in the case of *Tarzan of the Apes* does this meet the 95% ($p < 0.05$) confidence criterion.

6.2 The Constant Vocabulary Zipf-Bernoulli (CVZB) Model

Let $Q(\Pi_{ZB}, t)$ be a sample of t tokens generated by model $\Pi_{ZB}(\alpha, \hat{V})$, where α is the Zipf coefficient and $\hat{V}$ the maximum available vocabulary. ($\hat{V}$ should not be confused with the corpus vocabulary V.) Each successive token w is selected from a "dictionary" of types $\{w_1, w_2 \ldots w_{\hat{V}}\}$, where w_1 is the most probable and $w_{\hat{V}}$ the least probable; the subscript may be called the "rank", though this does not always correspond to the rank of the same word within $Q(\Pi_{ZB}, t)$. Under Zipf's law

$$p_{\Pi_{ZB}}(r) = Pr(w = w_r|\Pi_{ZB}) = \frac{1}{H_{\alpha,\hat{V}} \cdot r^{\alpha}} \tag{5}$$

where the normalizing constant $H_{\alpha,\hat{V}} = \sum_{i=1}^{\hat{V}} 1/i^{\alpha}$. Note that if $\hat{V}$ were infinite, the series would only converge if $\alpha > 1$ and $H_{\alpha,\hat{V}}$ would become the Riemann zeta function $\zeta(\alpha)$; however, with a finite positive $\hat{V}$ there is no restriction other than that $\alpha > 0$.

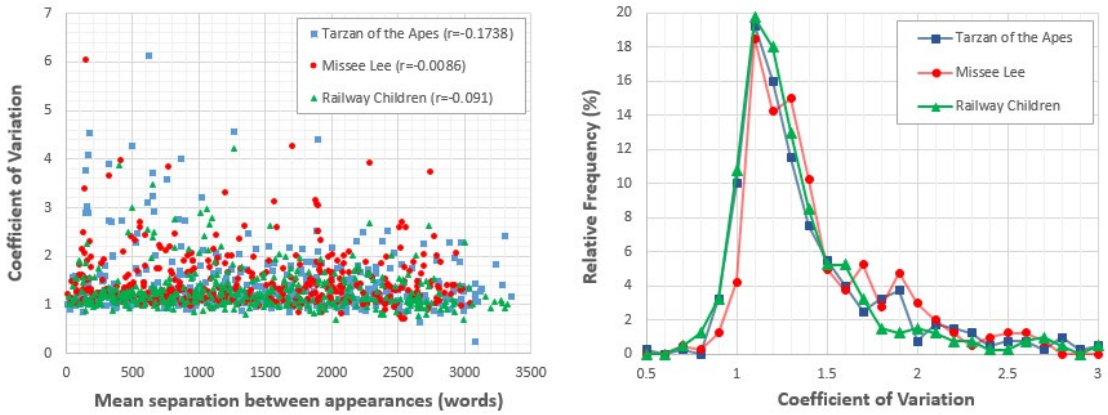

Figure 11: Coefficients of variations for word separations of the 400 most frequent words in three novels, showing a weak negative correlation with the mean separation. The modal value is approximately 1.1.

Now $v_{\Pi_{ZB}}(t) = \sum_{r=1}^{\hat{V}} \Lambda_{\Pi_{ZB}}(r, t)$ where $\Lambda_{\Pi_{ZB}}(r, t) = \begin{cases} 1; & w_r \in Q(\Pi_{ZB}, t) \\ 0; & w_r \notin Q(\Pi_{ZB}, t) \end{cases}$, whose expectation is the probability that at least one instance of w_r appears within $P(\Pi_{ZB}, t)$. Simple probability theory requires that $E[\Lambda_{\Pi_{ZB}}(r, t)] = 1 - [1 - p_{\Pi_{ZB}}(r)]^t$ (one minus the probability that w does *not* appear in the sample) so replacing $\Lambda_{\Pi_{ZB}}(r, t)$ with its expected value:

$$v_{\Pi_{ZB}}(t) \approx \sum_{r=1}^{\hat{V}} E[\Lambda_{\Pi_{ZB}}(r, t)] = \hat{V} - \sum_{r=1}^{\hat{V}} \left[1 - \frac{1}{H_{\alpha,\hat{V}} \cdot r^{\alpha}}\right]^t \tag{6}$$

where the latter summation term clearly represents the "unseen species" in the first t words. (This is equivalent to an expression in Boystov (2017), with the exception that in our case the limit of the summation is finite.) However, this model is not amenable to our optimization procedure (Section 3.2) since $\hat{V}$ is a discrete variable. To allow $\hat{V}$ to take non-integer values we define

$$v_{\Pi_{ZB}}(t) = \hat{V} - \sum_{r=1}^{\lfloor \hat{V} \rfloor} \left[1 - \frac{1}{H_{\alpha,\hat{V}} \cdot r^{\alpha}}\right]^t - (\hat{V} - \lfloor \hat{V} \rfloor) \left[1 - \frac{1}{H_{\alpha,\hat{V}} \cdot \hat{V}^{\alpha}}\right]^t \tag{7}$$

where $H_{\alpha,\hat{V}} = \sum_{i=1}^{\lfloor\hat{V}\rfloor} 1/i^{\alpha} + \frac{1}{1-\alpha}\left(\hat{V}^{1-\alpha} - \lfloor\hat{V}\rfloor^{1-\alpha}\right)$. We refer to this as the "Constant Vocabulary Zipf-Bernoulli" (CVZB) model. Applying the optimization procedure of Section 4.2, we find the optimal solution is independent of the initial assumption: however, for consistency all reported data were obtained using a starting hypothesis $\Pi_{ZB}(1.5, 10000)$.

Figure 12 shows the mean optimization error and the cumulative error frequency distribution for CVZB for the entire dataset, compared with the corresponding results of the Heaps and GT models. While CVZB is in general superior to either, it only beats the Heaps model by a narrow margin. The optimized parameters may now be compared with values determined independently from the same corpora. Figure 13 compares the optimised probability distribution associated with the CVZB model with relative word frequency distribution for *The Railway Children*. (Relative frequencies were normalized such that their sum equalled the approximate "coverage", i.e. the probability that the next token would be one previously encountered, estimated by Good (1953) to be $\left(1 - \frac{g_1}{T}\right)$. The slopes compare well in the moderately low frequency region (ranks 100-400) although the frequencies themselves only agree well at the extreme right of the graph.

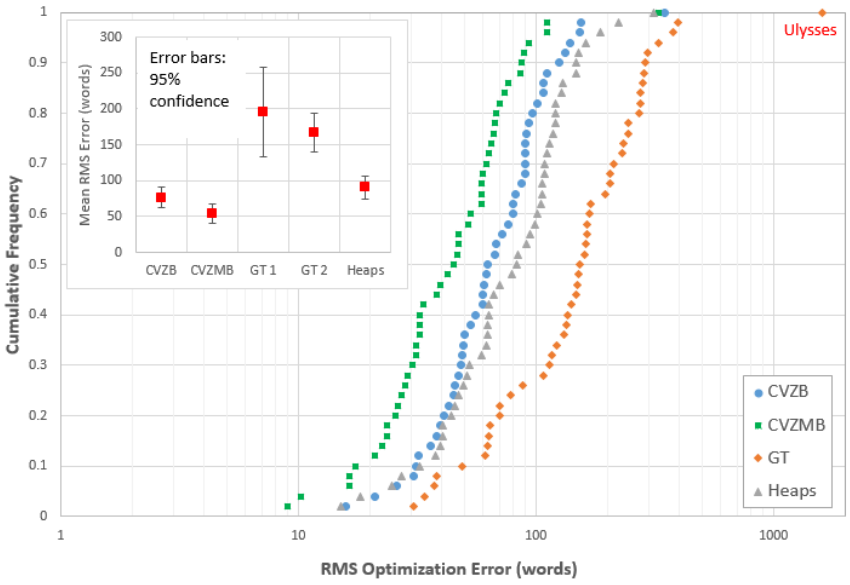

Figure 12: CVZB and CVZMB optimization error distributions for the entire dataset, compared with those of the GT and Heaps models. The inset compares the mean values for all four models. The GT mean of 195.56 (GT1) drops to 166.96 (GT2) when the outlier *Ulysses* is eliminated.

Figure 14(a) shows results for the entire dataset, plotting the three Zipf α coefficients obtained from the frequency distribution data against the optimized values for the same corpora; the latter correlates strongly with the "ultra low-frequency" indices computed indirectly using $\alpha = 1/(\beta - 1)$.

While the available vocabulary $\hat{V}$ obviously cannot be measured independently, one might intuitively expect it to be related to the vocabulary "richness" as quantified by the parameter ρ (see Figure 3). Figure 14(b) shows that such a strong correlation does indeed exist, with and without the two most egregious outliers (*Ulysses* and *The Book of Mormon*).

To test the predictive quality of the CVZB model, the latter was trained on the first 50% of each corpus and used to predict the final vocabulary. (The procedure was the same as that employed in Section 4.2.) Figure 15 shows the results, indicating that CVZB does in general produce a very significant underestimation of the total vocabulary.

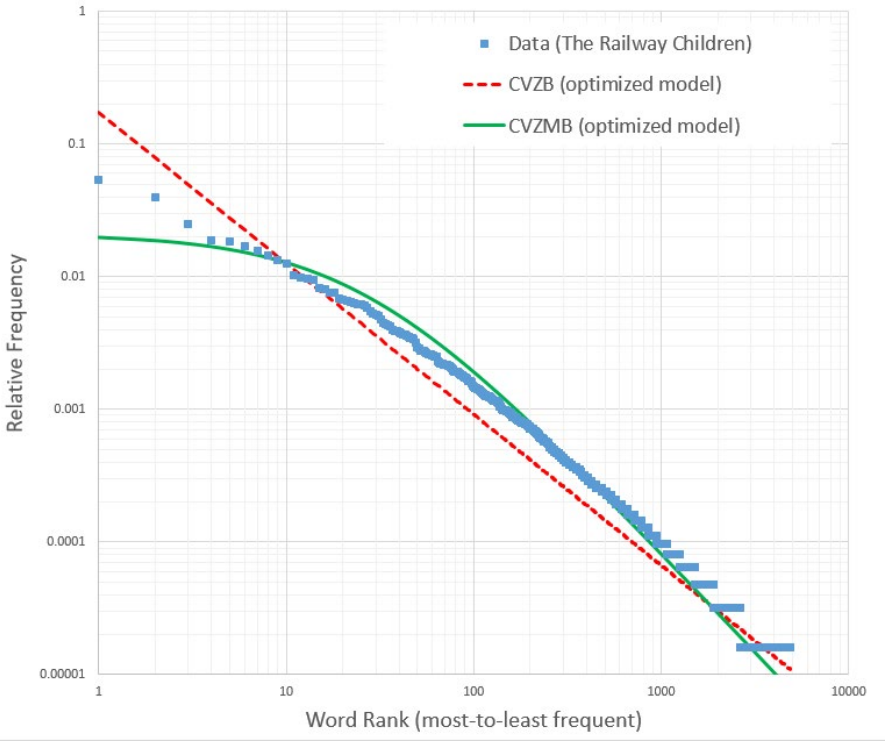

Figure 13: Normalized relative word frequency distribution for a typical text, compared with distributions obtained from optimized CVZB and CVZMB models.

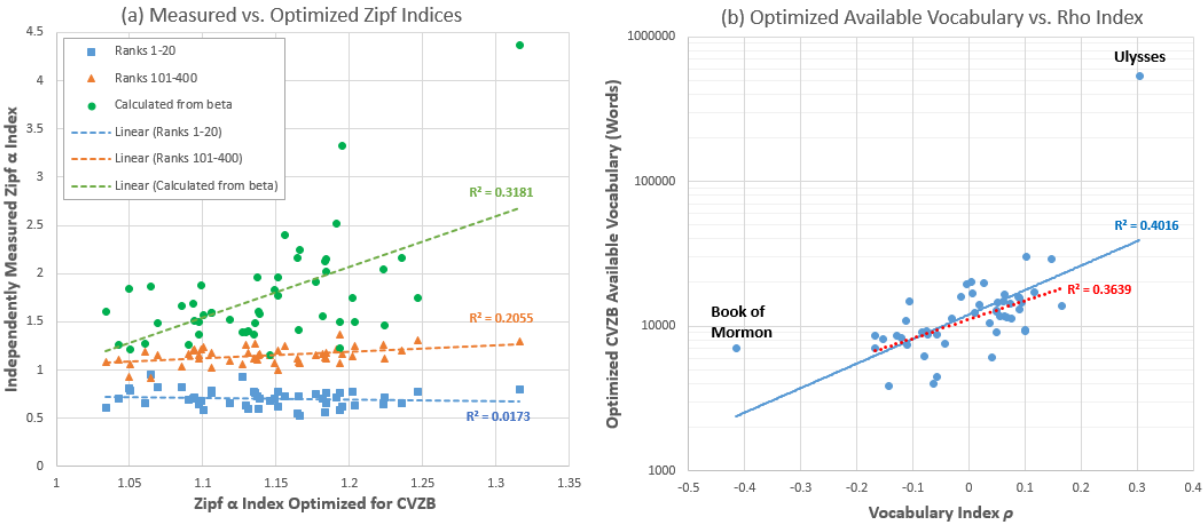

Figure 14: Correlation of parameters from optimized CVZB model with those independently measured. (a) The ultra-low frequency α indices (calculated indirectly from β) correlate quite strongly with the optimized values ($p = 0.00002$). The α indices for ranks 101-400 correlate less well (but still highly significantly with $p = 0.001$) while the high frequency indices show no significant correlation ($p = 0.363$). (b) A very strong correlation ($p < 0.00001$) exists between the optimized vocabulary size $\hat{V}$ and the logarithm of the independently measured vocabulary index ρ. (The red broken line indicates the trend with the two outliers removed.)

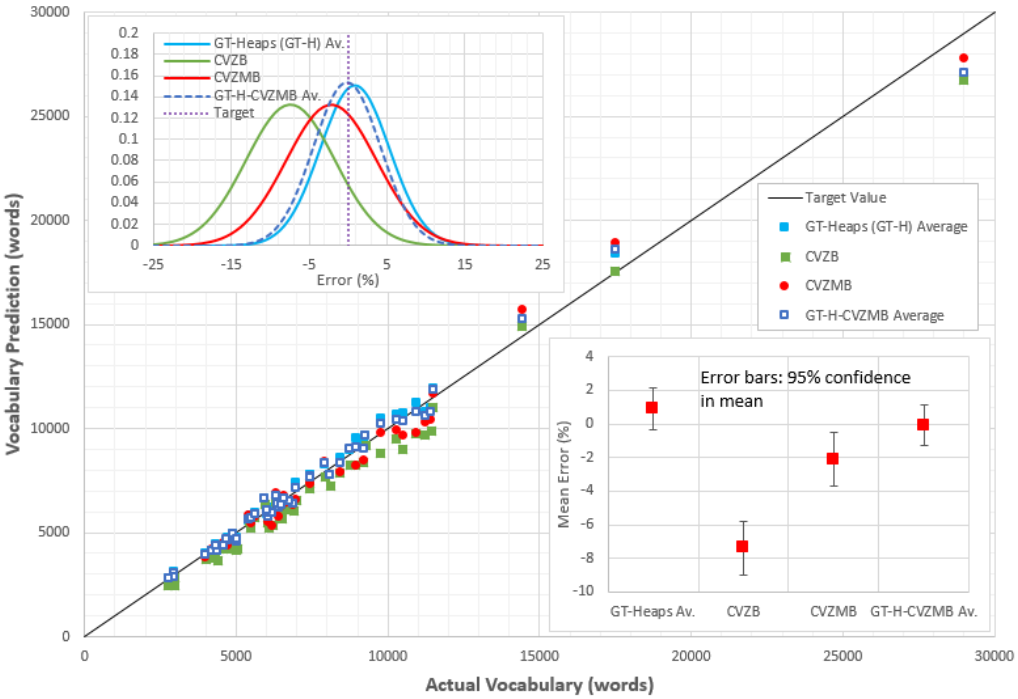

Figure 15: Vocabularies predicted from the first 50% of each corpus using the GT-Heaps average, CVZM and CVZMB models. Top left inset shows optimized Gaussians based on the mean and standard deviation errors of the two models. Bottom right inset shows the corresponding 95% error bars indicating that the average of GT, Heaps and CVZMB provides an extremely plausible unbiased estimator.

6.3 The Constant Vocabulary Zipf-Mandelbrot-Bernoulli (CVZMB) Model

With a view to obtaining a better prediction, we now modify the CVZB model to incorporate the Mandelbrot parameter, i.e. $\Pi_{ZMB}(\alpha, \hat{V}, m)$, meaning that three parameters must now be optimized. The equation for the vocabulary curve now becomes:

$$v_{\Pi_{ZMB}}(t) = \hat{V} - \sum_{r=1}^{\lfloor \hat{V} \rfloor} \left[1 - \frac{1}{H_{\alpha,\hat{V},m}(r+m)^{\alpha}}\right]^{t} - (\hat{V} - \lfloor \hat{V} \rfloor)\left[1 - \frac{1}{H_{\alpha,\hat{V},m}(\hat{V}+m)^{\alpha}}\right]^{t} \quad (8)$$

where $H_{\alpha,\hat{V},m} = \sum_{i=1}^{\lfloor \hat{V} \rfloor} 1/(i+m)^{\alpha} + \frac{1}{1-\alpha}\left(\left[\hat{V}+m\right]^{1-\alpha} - \left[\lfloor \hat{V} \rfloor + m\right]^{1-\alpha}\right)$. We refer to this as the Constant Vocabulary Zipf-Mandelbrot-Bernoulli (CVZMB) model.

The initial model used for optimization was $\Pi_{ZMB}(1.5, 10000, 10)$, though as before the final optimized model was found not to depend significantly on the original assumption. Unfortunately the model failed to meet the termination criterion within an acceptable timespan for corpus #47 (James Joyce's *Ulysses*), and the results for this corpus are based on 851 beneficial mutations of the original assumption. The RMS optimized error distributions are shown on Figure 12, showing a clear improvement over the CVZB model; however, this is only to be expected given the additional degree of freedom.

The vocabulary predictions for the CVZMB model are shown in Figure 15: while there is still some under-prediction, it is much less than it was for the CVZB model, though it is still inferior to the GT-Heaps average. Interestingly though, a simple average of GT, Heaps and CVZMB produces a *very* plausible unbiased estimator.

The word probability distribution computed for the CVZMB model optimized for a typical text is compared with the corresponding actual word frequency distribution in Figure 13, showing a better qualitative agreement than the CVZB model. However, although CVZMB produces lower optimization and prediction errors than CVZB, its optimized parameters do not correlate so well with those independently measured: see Figure 16, for which the correlation coefficients are summarized in Figure 17.

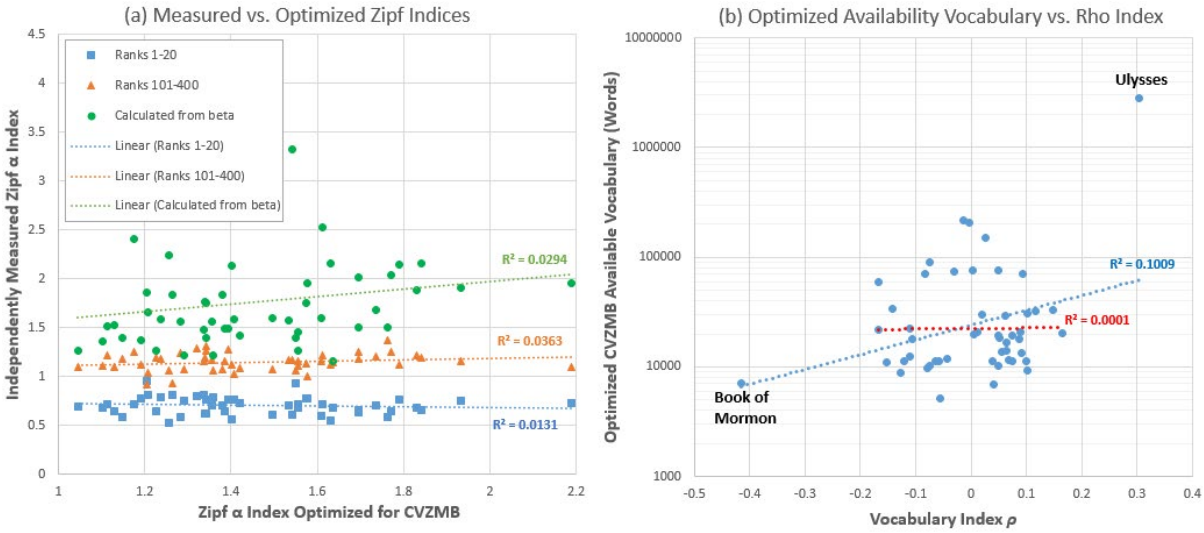

Figure 16: Correlation of parameters from optimized CVZMB model with those independently measured. (a) The α indices (calculated indirectly from β) no longer correlate significantly with the optimized values ($p = 0.274, 0.185$, and 0.429) as they did in the CVZB case (see Figure 14). (b) A weakly significant correlation ($p = 0.025$) exists between the optimized available vocabulary $\hat{V}$ and the logarithm of the independently measured vocabulary index ρ (blue broken line). However, this is largely due to two extreme outliers: with these removed (red broken line) the p-value becomes 0.945, indicating no correlation at all.

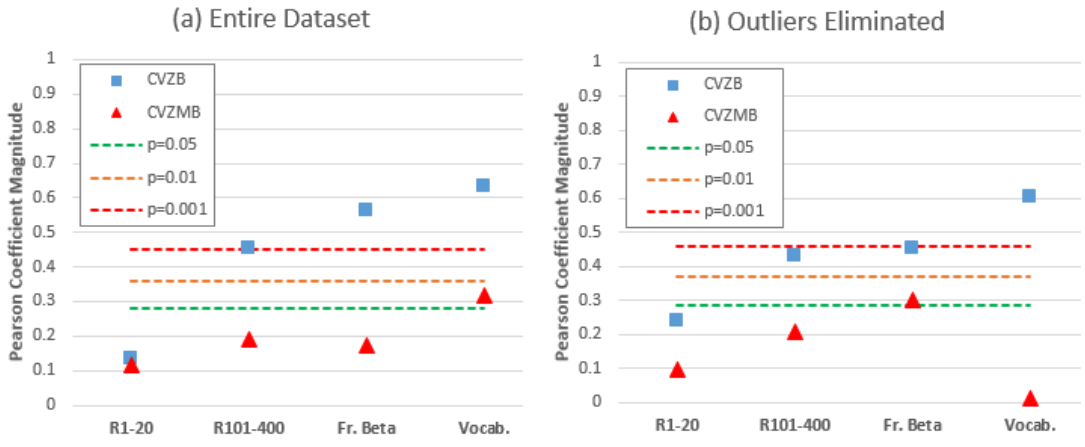

Figure 17: Pearson correlation coefficients and p-values for correlations between measured and optimized Zipf coefficients, and the measured vocabulary coefficient and optimized maximum vocabulary for CVZB and CVZMB models, with and without the outliers *Ulysses* and *The Book of Mormon*. (R1-20 and R101-400 refer to the rankings of the words considered. "Fr. Beta" means "calculated from β")

7. Conclusions and Future Work

The major observations of this paper are as follows.

1. The error obtained using the optimum Heaps' law model is significantly lower than that from the GT model.
2. "Future vocabulary growth" (i.e. the number of unseen types in a sub-sample revealed by extending the sample) is over-predicted by Heaps and under-predicted by GT; the average of the two provides a plausible unbiased estimator.
3. Although the data are widely scattered, it seems plausible that the optimized Heap indices are related to the Zipf α index in the ultra-low frequency domain (estimated using $\alpha = 1/(\beta - 1)$) by $\lambda = 1/\alpha$.
4. A model based on Bernoulli trial word selection with a Zipf distribution and a constant finite available vocabulary (CVZM) may be optimized to give results plausibly close to the data, with an optimization error marginally lower than that of Heaps.
5. The Zipf index optimized using CVZB correlated strongly with that calculated independently using $\alpha = 1/(\beta - 1)$.
6. The optimized maximum vocabulary for CVZB correlated strongly with the "vocabulary index" defined in (1).
7. CVZB significantly under-predicted future vocabulary growth.
8. The addition of the Mandelbrot generalization to the model (CVZMB) gave an even lower optimization error than CVZB.
9. The CVZMB model still under-predicted future vocabulary growth, but to a lesser extent than CVZB.
10. An equally weighted average of CVZMB, Heaps and GT provided (for the data-set tested) an almost perfectly unbiased estimate of future vocabulary growth.
11. The coefficients of correlation between measured Zipf indices and those optimized from CVZMB were almost zero.
12. With the two most egregious outliers removed, the maximum vocabulary optimized for CVZMB was not correlated at all with the vocabulary index.

While points 5 and 6 suggest an "element of truth" in the CVZB model which the CVZMB model lacks (cf. Points 11 and 12), the latter does provide a better prediction of the unseen types (Point 9).

The fact that averaging CVZMB, Heaps and GT provides an unbiased prediction is interesting but needs to be confirmed by further experimentation. In particular, the range of test data needs to be expanded: while we believe that the analysed texts do form a reasonable sample of typical English, a number of standardized corpora are also available for testing statistical language models. These include:

- The Brown Corpus (https://www.sketchengine.eu/brown-corpus/)
- The British National Corpus (http://www.natcorp.ox.ac.uk/)
- The Oxford English Corpus (https://www.sketchengine.eu/oxford-english-corpus/)

An important aspect of our future work will be to apply the same models to some of these standard corpora to determine whether the same results are observed.

Probably the most questionable aspect of both the CVZB and CVZMB models is the assumption that there exists throughout each corpus a static token probability distribution and a constant available vocabulary $\hat{V}$. If this were genuinely the case, one would expect $\hat{V}_{50}$ and $\hat{V}_{100}$ (the optimized values of V for the first 50% and 100% of each corpus respectively) to be approximately equal. Figure 18 shows

that these quantities do have a strong log-log correlation for both models ($p < 0.00001$ for CVZB and $p = 0.00011$ for CVZMB), but closer examination reveals an average increase in $\hat{V}$ as the corpus-size is extended. Figure 19 shows the cumulative frequency distributions of the "vocabulary increase ratio" $\hat{V}_{100}/\hat{V}_{50}$, showing that $E[\hat{V}_{100}/\hat{V}_{50}] \approx 1.5$ for CVZB and ≈ 2 for CVZMB, suggesting that on average the pool of available words grows as the corpus expands. However, this seems intuitively unreasonable. (Why should an author be selecting from a larger dictionary of words at the end of a book than at the beginning?) A more likely possibility is that the token-pool at any given instant is finite, and that its composition changes over time. (This would correspond to thematic variation in the plot of a novel, or the shifting of species population densities during a biometric survey.) The recorded vocabulary $v(t)$ however represents the aggregate over all these ephemeral token-pools. We would expect V optimised using $t \in [1, T]$ to exceed that optimised using $t \in [1, T/2]$ since in the former case more token-pools would have contributed. Further experiments will be needed to isolate these hypothesized transitory word-pools at different points during processing, and study their dynamic behaviour.

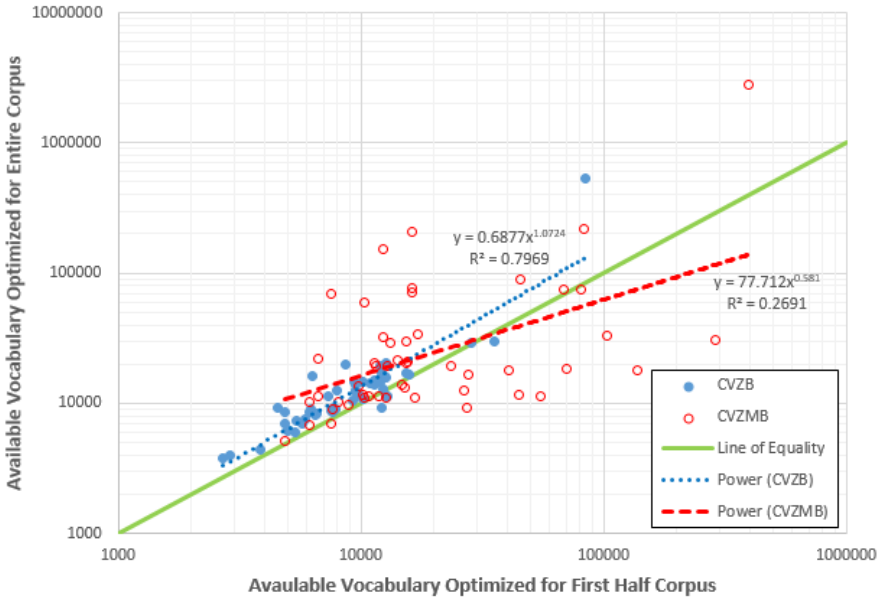

Figure 18: Comparison of the optimized available vocabularies for CVZB and CVZMB models using the first 50% and 100% corpus. (Correlation coefficients and optimal parameter values were obtained by linear regression in the log-log domain.)

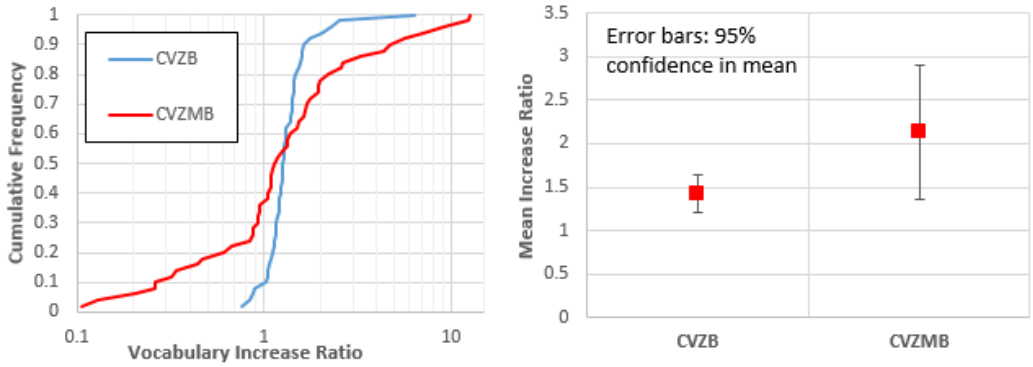

Figure 19: Left: the cumulative frequency distributions for the increase in optimised vocabulary ($\hat{V}_{100}/\hat{V}_{50}$) when the corpus used to train the model is doubled. Right: the mean values of $\hat{V}_{100}/\hat{V}_{50}$ showing 95% mean error bars. Interestingly, for the CVZB model, in 92% of cases $\hat{V}_{100} > \hat{V}_{50}$, but this falls to just 64% for CVZMB.

References

Altmann, E.G., Pierrehumbert, J.B., Adilson, E.M., 2009. Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distribution of Words. PLoS ONE, 4(11), e7678.

Bebber, D.P., Marriott, H.C., Gaston, K.J., Harris, S.A., Scotland, R.W., 2007. Predicting Unknown Species Numbers using Discovery Curves. Proc. R. Soc. B, 274, pp.1651-8.

Boystov, L, 2017. A Simple Derivation of the Heaps' Law from the Generalized Zipf's Law. arXiv:1711.03066v1.

Chiu, D-M, Jain, R, 1989. Analysis of increase and decrease algorithms for congestion avoidance in computer networks. Computer Networks and ISDN Systems, 17, pp.1-14.

Costello, M.J., Wilson, S, Houlding, B, 2012. Predicting Total Global Species Richness using Rates of Species Description and Estimates of Taxonomic Effort. Sys. Biol., 61(5), pp.871-83.

Dodd, J, Letts, P, 2017. Types, Tokens, and Talk about Musical Works. J. Aesthetics and Art Criticism, 75(3), pp.249-63.

Efron, B, Thisted, R, 1976. Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know. Biometrika, 63(3), pp.435-47.

Egge,L, 2007. Untangling Herdan's Law and Heaps' Law: Mathematical and Informetric Arguments. J. Assoc. Inf. Sci. Technol., 58(5), pp.702-9.

Fisher, R.A., Corbet, A.S. Williams, C.B., 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(1), pp.42-58.

Font-Clos, F, Corral, Á, 2015. Log-log Convexity of Type-Token Growth in Zipf's Systems. Phys. Rev. Lett., 114(23), 238701.

Good, I.J., 1953. The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika, 40(3/4), pp.237-64.

Good, I.J., Toulmin, G.H., 1956. The Number of New Species, and the Increase in Population Coverage, when a Sample is Increased. Biometrika, 43(1/2), pp.45-63.

Huang, X., Baker, J., Reddy, R., 2014. A Historical Perspective of Speech Recognition. Communications of the ACM, 57(1), pp 94-103.

Hunter, G., 2004. Statistical Language Modelling of Dialogue Material in the British National Corpus. PhD Thesis, University of London, U.K.

Hunter, G., Huckvale, M., 2006. Is it Appropriate to Model Dialogue in the Same Way as Text?. Proceedings of the European Modelling Symposium, London, U.K., September 2006, pp 199 – 203.

Hussain, K.S., 2015. Measuring Lexical Richness through Type-Token Curve: a Corpus-Based Analysis of Arabic and English Texts. Research on Humanities and Social Sciences, 5(4), pp.97-104.

Lü, L, Zhang, Z-K, Zhou, T, 2010. Zipf's Law Leads to Heaps' Law: Analysing their Relation in Finite-Size Systems. PLoS ONE, 5(12), e14139.

Mandelbrot, B, 1953. An Informational Theory of the Statistical Structure of Language. in Willis Jackson (Ed.), Applications of Communication Theory. pp.486-802, Butterworths, London.

Moore, R.K., 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. Proceedings of EuroSpeech '03, Geneva, Switzerland, pp 2582 – 2584.

Moore, R.K., 2001. There's no Data Like More Data – but when will Enough be Enough? Proceedings of the Institute of Acoustics, 23(3) .

Mora,C, Tittensor, D.P., Adl, S, Simpson, A.G.B., Worm, B, 2011. How Many Species are there on Earth and in the Ocean?. PLoS ONE, 9(8), e1001127.

Orlitsky, A, Suresh, A.T., Wu, Y, 2016. Optimal Prediction of the Number of Unseen Species. PNAS, 113(47), pp13283-8.

Richards, B, 1987. Type/Token Ratios: What do they Really Tell Us. J. Child Lang., 14, pp.201-9.

Rudman, J, 2000. Non-Traditional Authorship Attribution Studies: Ignus Fatuus or Rosetta Stone?. BSANZ Bulletin, 24(3), pp.1963-76.
van Leijenhorst, D.C., van der Weide, Th.P., 2005. A Formal Definition of Heaps' Law. Information Sciences, 170, pp.263-72.
Wetzel, L, 2018. Types and Tokens. The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Zalta, E.N. (Ed.), https://plato.stanford.edu/archives/fall2018/entries/types-tokens/.
Youmans, G, 1990, Measuring Lexical Style and Competence. Style, 24(5), pp.584-99.
Zipf, G.J., 1949, Human Behaviour and the Principle of Least Effort. Addison-Wesley.

Author Biographies

Martin Tunnicliffe studied Electrical and Electronic Engineering at the University of Bradford, UK, and undertook postgraduate research in Semiconductor Failure Mechanisms at Loughborough University, UK. He obtained his PhD in 1993. After a postdoctoral fellowship in the Loughborough High Speed Networks Group, he became a Lecturer at Kingston University in 1997. His current research interests include Network Simulation and Modelling, Cyber Security, Information Theory and Type-Token Systems.

Gordon Hunter obtained his BA in Mathematics at Churchill College, University of Cambridge. After several years teaching Mathematics and Physics in colleges, he did postgraduate studies at University College London, where he completed an MSc in Electronic Engineering and Computer Science in 1999 and a PhD in Computer Speech Technology in 2004. He moved to Kingston University in 2003, and is now Senior Lecturer in Mathematics and Computing. His research interests include Speech, Language and Acoustic signal processing, Statistical Machine Learning and various applications of Mathematics and Statistics to problems in Science, Medicine and Education.