entropy

MDPI

Article

Entropy Estimation Using a Linguistic Zipf–Mandelbrot–Li Model for Natural Sequences

Andrew D. Back * and **Janet Wiles**

School of Information Technology and Electrical Engineering, The University of Queensland,
Brisbane, QLD 4072, Australia
* Correspondence: a.back@uq.edu.au; Tel.: +61-7-3365-1111

Abstract: Entropy estimation faces numerous challenges when applied to various real-world problems. Our interest is in divergence and entropy estimation algorithms which are capable of rapid estimation for natural sequence data such as human and synthetic languages. This typically requires a large amount of data; however, we propose a new approach which is based on a new rank-based analytic Zipf–Mandelbrot–Li probabilistic model. Unlike previous approaches, which do not consider the nature of the probability distribution in relation to language; here, we introduce a novel analytic Zipfian model which includes linguistic constraints. This provides more accurate distributions for natural sequences such as natural or synthetic emergent languages. Results are given which indicates the performance of the proposed ZML model. We derive an entropy estimation method which incorporates the linguistic constraint-based Zipf–Mandelbrot–Li into a new non-equiprobable coincidence counting algorithm which is shown to be effective for tasks such as entropy rate estimation with limited data.

Keywords: entropy estimation; Zipf–Mandelbrot–Li law; language models; probabilistic natural sequences

check for updates

Citation: Back, A.D.; Wiles, J. Entropy Estimation Using a Linguistic Zipf–Mandelbrot–Li Model for Natural Sequences. *Entropy* **2021**, *23*, 1100. https://doi.org/10.3390/e23091100

Academic Editors: Sergio Cruces, Iván Durán-Díaz, Rubén Martín-Clemente and Andrzej Cichocki

Received: 15 July 2021
Accepted: 19 August 2021
Published: 24 August 2021

1. Introduction

Natural systems such as language, can be understood in terms of symbolic sequences described within an information-theoretic framework, where meaning is encoded through the arrangement of probabilistic elements. When placed in a mathematical framework, we can characterize and begin to understand the meaning of messages, not only on the basis of the meaning directly attached to words, but on the statistical characteristics of symbols.

Using this approach, natural language can be viewed as observing one or more discrete random variables X of a sequence $X = X_1, \ldots, X_i, \ldots, X_K$, $X_i = x \in \mathbf{X}^M$, that is, x_i may take on one of M distinct values, $\mathbf{X}^M$ is a set from which the members of the sequence are drawn, and hence x_i is in this sense symbolic, where each value occurs with the probability $p(x_i), i \in [1, M]$.

The single symbol Shannon entropy (if not otherwise specified, any reference to entropy will refer to the classical Shannon entropy of unigram probabilities.) is defined for unigrams as [1,2]

$$H_0(X) = -\sum_{i=1}^{M} p(x_i) \log_2(p(x_i)) \tag{1}$$

In the context of language processing, statistical models of symbol sequences are of interest and can be defined by the probability $p(\Omega) = p(s_1, \ldots, s_N)$ where Ω is a sequence of N symbols $\{s_i\}$. If the full sequence is available, the n-gram entropy can be directly computed using the same formula as (1), where instead of computing unigram probabilities, the joint probabilities are estimated so the n-gram entropy is obtained. However, the problem with this approach is that a large amount of data is generally required and the reliability is questionable for $N > 5$ [3].

In probabilistic language modeling, it is often of interest to predict the next symbol in a sequence using the previous symbols. Hence, the joint probability can be computed using the Markov property by considering the previous block of symbols in terms of the conditional probabilities as

$$p(s) = \prod_{i=1...N} p(s_i|s_1 \ldots, s_{i-1}) \tag{2}$$

Hence, provided the symbol probabilities can be estimated, it is possible to determine the n-gram probabilities and n-gram entropy of language sequences.

A related method of characterizing language models is to measure perplexity [4], defined as

$$P_e(s) = 2^{H(s)} \tag{3}$$

In contrast to entropy, which can be understood as measuring the average number of bits to encode the information in a symbol, perplexity can be intuitively understood as measuring the total number of bits required to encode the information in a sequence; hence, the smaller the value the better.

While entropy provides a measure of information in a given sequence, this raises the question of how quickly the information grows with increasing text length [1,5]. The idea is that the entropy rate can measure the complexity of language by the average information content of symbols such as words taken over a sufficiently long period. Similarly, the effectiveness of compression algorithms can be measured by how closely the algorithm can compress any stationary and ergodic source down to the entropy rate for a sufficiently long input source sequence [6].

Entropy rate has been of interest for analyzing the information content neuronal spike sequences [7], complexity of short heart period variability [8], attention models using visual salience attention [9], complexity of animal vocal complexity [10], statistical structure of non-redundant coding sequences in DNA [11], behavior prediction [12] and in estimating the long-term memory of language models [13].

A problem with entropy estimation is characterizing infrequently occurring symbols, hence requiring a potentially large number of samples to adequately model the probabilistic structure [14].

In contrast to statistical descriptive applications which depend on large amounts of data, we are interested in building models of social interaction using entropy estimation methods with limited available data.

Various efficient methods of entropy estimation have been proposed. Entropy estimation over short symbolic sequences for dynamical time series models was considered in [15]. A computationally efficient method for calculating entropy based on a James–Stein-type shrinkage estimator was proposed in [16]. Methods for overcoming bias in maximum likelihood entropy estimators with limited data have been examined in [17]. The Nemenman–Shafee–Bialek (NSB) entropy estimator extends this concept to correct sample size-dependent bias by using a Bayesian approach to construct priors with power–law dependence on the probabilities, in particular, using Dirichlet distributions [18].

The advantages of more sophisticated entropy estimation techniques which go beyond naive plug-in methods are evident. These can be broadly referred to as "model-based" because they introduce some additional complexities into an otherwise simple algorithm which takes into account some understanding of the nature of the data and the estimation process whilst remaining broadly applicable. For example, a model-based estimator using an understanding of how sequences of symbols will have probabilistic patterns of "coincidence" was proposed in [19].

In this paper, we consider a novel probabilistic model-based entropy estimator which extends [19] and is comparable to [18] in that it uses a limited amount of data and an a priori model as a basis for constructing an efficient entropy estimator. The model "hint" that we introduce is the idea that for many natural sequences including language, instead of a naive estimator, the probabilistic distribution of symbols is expected to follow linguistic patterns.

Hence, the basis for our proposed approach is to develop an analytic rank-based Zipfian-style probabilistic model which is constrained to accommodate the linguistic features of human language and to incorporate this into an efficient non-equiprobable coincidence counting the entropy estimation algorithm.

In the next section, we describe a coincidence-counting entropy estimator and introduced the concept of linguistic entropy. In Section 3, we introduce a new framework of linguistic probabilistic models which is incorporated into the proposed entropy estimator. In Section 4, we demonstrate the efficacy of the proposed model and show that it provides a high degree of accuracy while requiring a small number of samples.

2. Model-Based Entropy Estimation

2.1. Coincidence Counting Approach

Entropy estimation difficulties can occur with low probability events exacerbated due to real-world issues surrounding the data, including problems of small data sets [20], limited resource environments [21], bias due to heavy-tailed distributions [22] and uneven distributions with poorly populated bins [23]. The latter problem is especially evident in estimating entropy in language involving very low probability events such as infrequent words.

The problem of undersampling in the context of entropy estimation, where the alphabet size is large compared with the number of samples, that is, $N \ll M$, has been considered at length where it is well known that significant bias can occur, particularly in the case of using binning approaches [23,24]. Hence, alternative entropy estimation algorithms are of interest which can provide useful results with a small number of samples [25–27].

One class of proposed solutions is based on the method of coincidence counting to derive entropy from the phase space trajectory of symbolic events [28]. In particular, Ma proposed the method of coincidence counting as a suitable method of deriving entropy from the phase space trajectory, noting the problematic issue of metastability with estimating the empirical probability distribution. A simple algorithm for entropy estimation based on this approach was proposed in [19]. Their novel approach used the idea of estimating probabilities from a quadratic function of the inverse number of symbol coincidences; however, it has the limitation of this method, which was that it assumed equiprobable symbols. In the next section, we show how it is possible to extend this to the non-equiprobable case by using an analytic Zipf–Mandelbrot–Li law.

2.2. Linguistic Entropy Estimation

Consider a sequence of symbols which is defined by a discrete (or symbolic) random variable x which may take on a finite number M of distinct values $x_i \in \{x_1, \ldots, x_M\}$ with probabilities $p(x_i), i \in [1, M]$. For example, suppose $M = 4$ and we have a sequence of symbols *abcdabc*. The frequency of symbols can be estimated as a function of the distance between consecutive repeating symbols or the 'coincidence distance' and in this case, the initial distance for a is $D(a; M) = 5$. Hence, it can be observed that by measuring this distance, it may be possible to estimate the relative frequency of any given symbol by measuring the distance between them.

To compute the probability $f(n; M)$ of a first coincidence occurring exactly at the nth symbol for $1 < n < M$ means that it is necessary to compute the probability of drawing no repeating symbols in the entire sequence up to the $(n-1)$th draw given by $\widetilde{F}(n-1; M)$ and consequently drawing any $q_{n-1} \in [2, \ldots, n-1]$ identical symbols is given by $F(n-1; M)$. Hence, the nth symbol coincidence probability is given by [19]

$$f(n; M) = F(n; M) - F(n-1; M) \tag{4}$$

The expectation of the discrete parameter n and its associated probability $f(n; M)$ is given by

$$E[n] = J(n;M) \tag{5}$$
$$= \sum_{n=0}^{M} n f(n;M) \tag{6}$$

Since n is a function of M, we may define the coincidence distance:

$$D(M) = E[n]. \tag{7}$$

By forming a model to estimate M using the symbol distance D such as

$$\widehat{M}(D) = G(\Theta; D) \tag{8}$$
$$= \sum_{i=0}^{n_p} \theta_i D^i \tag{9}$$

then by measuring $D(M)$ and hence evaluating M from the parametric model, then the entropy can be directly estimated from the symbol coincidences.

The model parameters θ_i can be determined by fitting a curve to an ensemble of data. For equiprobable symbols, the Shannon entropy is estimated as

$$H_0(M) = \log_2(\widehat{M}(D)) \tag{10}$$

Using this approach with only a small number of symbol observations, entropy estimation for equiprobable symbols was shown to be accurate and with a low bias of [19,29].

Now, for any given M, each symbol of a specified rank r can be treated as being equiprobable and hence by considering the probability of each ranked symbol, then we have:

$$\widetilde{F}(n;M) = 1 \cdot (1 - P_2) \cdot (1 - P_3) \cdots (1 - P_{n-1}) \tag{11}$$

where $\widetilde{F}(n;M)$ is the probability of drawing any symbol on the first try followed by any other different symbol up to the nth draw and up to n symbols, and P_h is the probability of independently drawing $h-1$ identical symbols from a set of M in $h-1$ draws.

It follows that we can define the probabilities in terms of rank using a probabilistic model such as the Zipf–Mandelbrot–Li law developed by Li [30], who showed that the constants can be computed as

$$\alpha = \frac{\log_2(M+1)}{\log_2(M)} \tag{12}$$
$$\beta = \frac{M}{M+1} \tag{13}$$
$$\gamma = \frac{M^{\alpha-1}}{(M-1)^{\alpha}} \tag{14}$$

with a normalization step introduced in [14] as

$$\gamma' = \frac{\gamma}{\kappa} \tag{15}$$

to give:

$$\sum_{i=1}^{M} p(i) = 1, \quad \sum_{i=1}^{M} \frac{\gamma}{(r+\beta)^{\alpha}} = \kappa \tag{16}$$

which leads to:

$$P(r;M) = \frac{\widetilde{\gamma}}{\left(r(L) + \widetilde{\beta}\right)^{\widetilde{\alpha}}} \tag{17}$$

This approach provides an equiprobable representation of the symbols by considering a different model for each symbol rank.

Hence, an invertible rank-based model for $D(M) = J'(n, r, M)$ such as a power-based model is chosen so that the inverse model can be directly estimated using the forward model, for example, as

$$\widehat{D}(M;r) = \frac{1}{a}\left(\widehat{M}(r;D) - c\right)^{\frac{1}{b}} \tag{18}$$

where $\theta = \{a, b, c\}$ are the forward parameters.

Given an estimate $\widehat{M}(D;r)$ from the observed inter-symbol distance, it is possible to apply this parameter to the Zipf–Mandelbrot–Li set of equations in addition to our rank-based probability model, and estimate the entire set of symbolic probabilities. Using $\widehat{P}_h(r, M)$, the entropy can then be easily estimated as

$$\widehat{H}_1(r, X) = -\sum_{h=1}^{\widehat{M}} \widehat{P}_h(r, M)\log_2\left(\widehat{P}_h(r, M)\right) \tag{19}$$

which defines the rank r Shannon entropy estimate.

The model is applied by determining the mean distance between symbols $D_i(r)$ and then finding the estimated value $\widehat{M}(D;r)$ which is used to estimate entropy. The scaling of the inverse model curves for the proposed algorithm can be observed in Figure 1.

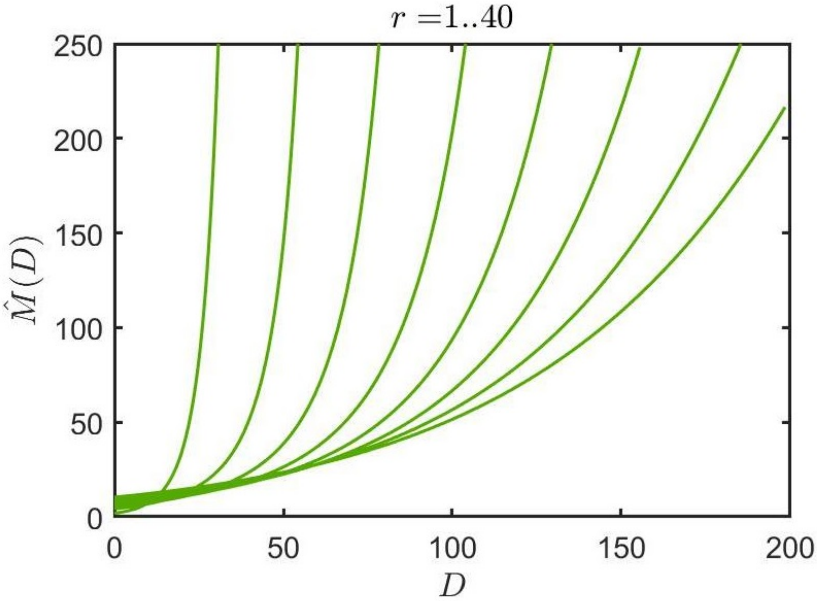

Figure 1. The inverse model curves (1..40) for the proposed algorithm are shown here for the range of the top 40 ranked symbols, also shown here for an alphabet size of $M = 200$.

2.3. Remarks on Bias and Convergence Properties

The proposed algorithm is defined in terms of a Zipf–Mandelbrot–Li distribution which uses a coincidence counting approach to estimate the mean symbolic distance between one or more ranked symbols and then use this to form an estimate of the whole distribution. We can consider the bias and convergence properties in terms of the maximum likelihood estimator for $\widehat{D}(M;r)$ in contrast to the probabilities directly in a plug-in entropy estimator.

Algorithmic bias can occur due to systematically underestimating the mean distance between symbols $D_i(r)$ [29]. To compute the bias precisely requires a closed form of the probability density function of $D_i(r)$. Analyzing this requires considering (4) and (11)–(16), where an approximation to the probability distribution can be obtained from the multiplicative process defined by (11) in terms of a log-normal distribution [31]. However, since a closed form solution for the likelihood of a log-normal distribution is not generally available, it is non-trivial to determine the specific bias properties [32,33]. One possible solution to this is to examine the probabilistic bounds [34]. Though we do not consider it in this paper, the bias properties of the proposed algorithm may also be improved by applying techniques such as the Miller–Madow procedure [16].

In terms of the convergence properties of the algorithm, a full proof of convergence properties is beyond the scope of this paper; however we provide an indication of some properties of the expected result. Note that $D_i(r)$ can be determined from the maximum likelihood estimation of the inter-symbol distance for any given symbol. Choosing the most frequent symbol $r = 1$, then $D_i(1)$ will converge in the sense of a usual maximum likelihood estimator to within some limits with a particular confidence level [35].

Convergence depends on symbols with rank $r = 1$, with specified probability $\widehat{P}_h(1, M)$ and all other possible symbols with probability $1 - \widehat{P}_h(1, M)$. Hence, the number of symbols required to estimate $D_i(r)$ to within the specified degree can be found by means of the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [35].

Since the convergence of $D_i(r)$ depends on the estimation of $\widehat{P}_h(r, M)$, then the DKW inequality provides the following result [14]:

$$P\left\{\sup_{r\in\mathbb{N}}\left|\widehat{P}_h(r,M) - P(r,M)\right| < \epsilon_r\right\} \leq \zeta' \tag{20}$$

where for a maximum difference ϵ_r between the estimated probability $\widehat{P}_h(r, M)$ and its theoretical target value $P(r, M)$, there will be n_r samples required to estimate the probability with a confidence level of ζ', specified by

$$\zeta' = 1 - 2e^{-2n_r\epsilon_r^2} \tag{21}$$

Hence, following [14], it can be shown that for a given confidence level, the minimum number of samples required to estimate $P(r, M)$ which are described by a Zipf–Mandelbrot–Li approximation, can be found as

$$N_r \leq \frac{8}{P(r,M)\Delta_r^2}\ln\left(\frac{2}{1-\zeta'}\right) \tag{22}$$

where for a ranked distribution, Δ_r is found as

$$\Delta_r = P(r,M) - P(r+1,M) \tag{23}$$

Now, it follows that the relative convergence performance can be defined in terms of a scaling factor $\lambda_f(M)$ which measures the reduction in samples required for convergence compared to a naive plug-in estimator as measured against the symbolic alphabet size M.

Hence, we have:

$$\lambda_f(M) = \frac{\frac{8}{P(M,M)\Delta_M^2}\ln\left(\frac{2}{1-\zeta'}\right)}{\frac{8}{P(1,M)\Delta_1^2}\ln\left(\frac{2}{1-\zeta'}\right)} \tag{24}$$

which simplifies to:

$$\lambda_f(M) = \frac{P(1,M)\Delta_1^2}{P(M,M)\Delta_M^2} \tag{25}$$

A graph of the relative convergence performance is shown in Figure 2 where an improvement of several orders of magnitude in the reduction in the number of samples required can be observed for alphabet sizes in the ranges $M = 20 - 40$ which are of typical interest in both natural and synthetic languages.

In contrast, the conventional plug-in estimator requires an estimation of all symbol probabilities which depends on the probabilities of all symbols and consequently significantly more samples. A full derivation of this latter result is shown in [14]. Hence, the proposed entropy estimation algorithm converges with a factor of approximately $\lambda_f(M)$ fewer symbols than in the conventional case.

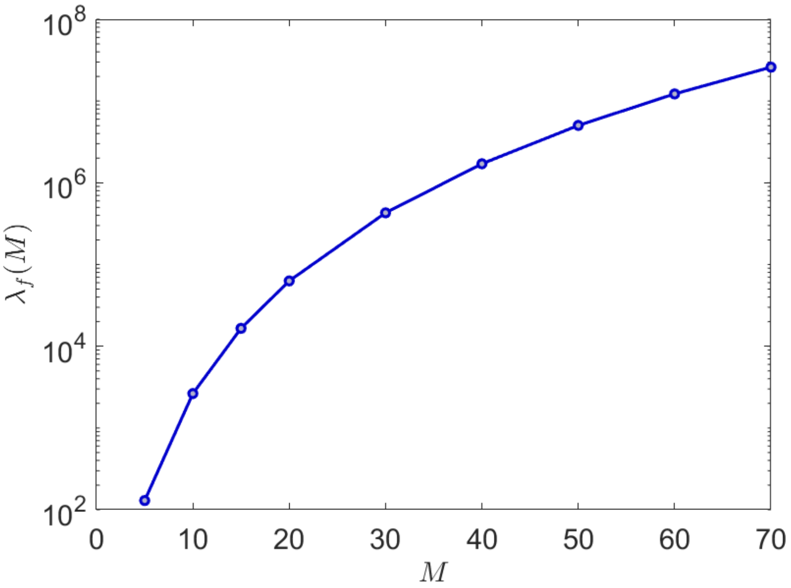

Figure 2. This graph shows the relative convergence performance $\lambda_f(M)$ of the proposed algorithm scaled against a conventional plug-in entropy estimation algorithm as measured against the symbolic alphabet size M. Note that the improvement is easily several orders of magnitude for alphabet sizes of interest in the range 20–40.

The current estimator employs a ZML distribution, and in the next section, we extend this model to include linguistic constraints.

3. Linguistic Probabilistic Models

3.1. Limitations of Zipfian Models for Language

For various natural sequences, Zipf's law describes how the frequency of ranked events occurring in such a way that they can be described by a power law [36–38]. This question of whether Zipf's law is a universal law of natural language and other phenomena has generated substantial interest over a long period of time [39].

The premise of Zipf in 1949 was that natural systems follow a principle of least effort, which means that individuals will follow a course of action which involves the expenditure of the least amount of work [40]. In terms of human language, this implies that the distribution of word use would follow the same principle so communication would occur efficiently with the least effort.

Various ongoing works have attempted to prove and disprove results in this field. Miller proposed that a monkey typing would produce a natural language with Zipfian distribution [41]. This argument was based on the result that the probabilistic distribution of words in natural languages only occurs as a statistical artifact of random spaces and can be described by Zipf's law.

While this claim has continued to generate considerable interest decades later, in fact, Miller's result was shown to be flawed by Howes in 1968 [42]. The problem with Miller's result is that assumes all word probabilities are strictly ranked by word length. Moreover, it assumes that all possible words of the same length have the same probability and that all sequences of letters are equiprobable. Clearly, these assumptions are not valid for natural language. A more recent analysis of this problem was performed, for example, considering unequal letter probabilities and log normal rank distributions [43,44]. It was found in [45] that the average information content was more consistently ranked than word length, by examining the inter-word statistical dependencies as the n-gram entropy of words in a local linguistic context.

Cancho and Sole considered the principle of least effort in human language as a compromise between speaker's and hearer's needs, where they were able to show results which indicate how Zipf's law explains the observations [46]. The concept of efficiency in languages was considered in [47], where efficiency can be defined in terms of successfully transmitting many different messages with minimal effort, yet balanced in terms of informativeness and complexity. Principles of least effort are closely related to the concept of semantic language universals by which such effort can be instantiated and measured. Accordingly, the principle of *ease of learning* was found to have strong evidence as a language universal in [48].

Zipf's law has been shown to occur as a result of the choice of rank as an independent variable [30,49], and hence has been challenged in terms of suitability as a universal model of human language or other natural sequences [50]. For example, in [30], it was reported that because the word frequency distribution of random texts can exhibit Zipfian characteristics, then Zipf's law is unsuitable as a criterion for identifying natural languages. However the statistical analysis of this result was disputed in [51] where it was shown that the rank distributions of random and natural texts are statistically inconsistent, and this suggests that Zipf's law may exist as a fundamental principle in natural languages.

While word frequency has generated significant interest, Zipf's law may operate at other levels in natural language, for example some results show the ranked order of phrases in natural language [52]. Moving towards more complex understandings of how Zipf's law can be refined, Corral considered the issue of how Zipf's law applies to normalized language element lemmas, i.e., a stem-like word form [50], and how a more complex formulation of Zipf's law of word frequency arising from a mixture of conditional distributions of frequency at different lengths may provide a better explanation of observations [53].

It is evident that there is not yet a single definitive answer as to the question of whether Zipf's law is necessarily a universal model of human language. However, it is clearly useful in forming a model of ranked symbolic information transmission which mimics human language elements [49,53] and have proved helpful as a probabilistic model for characterizing the observed behavior of natural symbolic sequences [54]. Here, we do not seek to prove the universality of Zipfian laws for language, but we consider their use as a way to model some aspects of natural language and how this may be useful for entropy estimation.

While a number of variations of Zipf's law have been proposed, including the Bradford Law [55] and Lotka Law [56], we previously proposed a new variation of the model we refer to as the Zipf–Mandelbrot–Li law [14,30,57–60], which models the frequency rank r of a language element $x \in \Sigma^{M+1}$ from an alphabet of size $M+1$, then, for any random word of length L, given by $v_k(L) = \{w_s, x_1, \ldots, x_L, w_s\}$, $k = 1, \ldots, M^L$ the frequency of occurrence is determined as

$$p_i(L) = \frac{\lambda}{(M+1)^{L+2}} \qquad i = 1, \ldots, M^L \tag{26}$$

where Li showed that λ can be analytically determined [30].

While these various rank-based laws provide a convenient analytical framework to model symbolic sequences, there is a problem in terms of known languages because such simple models carry no particular linguistic information. For example, consider the case of five-letter words. In English, there are approximately 12,500 known five-letter dictionary words. However, in the unconstrained ZML model, there can be 11.8 M words allowed. In the unconstrained form, this means that we allow words such as aaaag, rrrrx, czzzs, xyyaa and many other invalid words. A way to understand this is that such words do not conform to known linguistic principles such as orthographics [61], syllable structure [62], consonant/vowel ratio and organization [63,64], letter position [65] and graphemes [66]. Each of these can be viewed as a constraint on the allowable letters and their position in any word and hence it is evident that an unconstrained word-letter model allows a greater number of words than should be anticipated.

This raises the question of whether it is possible to introduce a probabilistic model which extends the advantageous Zipf–Mandelbrot–Li law to one which includes some linguistic constraints. Such a model would potentially provide the same useful analytic formulation while providing a more realistic bound on the types of words implicitly permitted. While the ZML model offers an effective basis for computing linguistic behavioral characteristics, it is evident that there is a need for improved models which provide a greater degree of conformity to the true properties of human language if such models are to be better utilized.

3.2. Unconstrained Rank-Ordered Probabilistic Model

Given a natural sequence such as language elements, consider a Zipfian probabilistic model of symbolic events which models the frequency rank r of a word (a word or n-gram is not necessarily referring to human language, but indicates a specific set of sequentially occurring symbols.), i.e., the r-th most frequent word, by a simple inverse power law, such that the frequency of a word $f(r)$ scales according to an equation which is given by

$$f(r) \propto \frac{1}{r^{\alpha}} \tag{27}$$

where a proportionality-dependent constant on the particular corpus may be introduced, Ref. [30] and where typically $\alpha \approx 1$. Thus, if $p_i(x)$ follows a Zipfian law, then $p_0(x) \propto 1/M$ and $p_i(x) = \varphi f(r)$. This power law equation can be understood in terms of the principle of least effort. It indicates that frequency decays linearly as the rank increases according to a log–log scale.

A way to view this is that according to Zipf's law, efficient language will minimize the effort between speaker and hearer [46]. Hence, the speaker may use a small vocabulary of common words to minimize effort in speaking and the hearer may desire a large vocabulary of less common words to minimize the effort in terms of ambiguity or confusion (Note that ambiguity and confusion are different concepts. The former may define the meaning of words, whereas confusion can relate to the intelligibility of words. Zipf's law provides a mathematical explanation of the balance between these competing features.).

We consider the Zipf–Mandelbrot law below [58]:

Given symbols $x \in \Sigma^{M+1}$ from an alphabet of size $M+1$ which includes a blank space w_s then for any random word of length L, given by $v_k(L) = \{w_s, x_1, \ldots, x_L, w_s\}$, $k = 1, \ldots, M^L$ the total number of words possible is given by

$$\begin{aligned} N_w &= \prod_{k=1}^{L} M_k, \quad k = 1, \ldots, L \\ &= M^L \end{aligned} \tag{28}$$

It follows that the frequency of occurrence for an unconstrained word of length L following a Zipf–Mandelbrot–Li distribution is determined as

$$p_i(L) = \frac{\gamma}{(M+1)^{L+2}} \qquad i = 1, \ldots, M^L \tag{29}$$

where γ is a normalization constant. Now, the summation of all probabilities of all such words is given by

$$\sum_{L=1}^{\infty} N_w(L) p_i(L) = 1 \tag{30}$$

$$= \sum_{L=1}^{\infty} \frac{\gamma M^L}{(M+1)^{L+2}} \tag{31}$$

$$= \frac{\gamma M}{(M+1)^2} \tag{32}$$

and hence the normalization constant can be found as

$$\gamma = \frac{(M+1)^2}{M} \tag{33}$$

Li subsequently showed that, using an exponential transformation from the word length to word rank model, it is possible to derive a rank ordered, parametric probabilistic model-which extends the Zipf–Mandelbrot model and is defined in terms of the alphabet size M [30].

3.3. Constrained Linguistic Probabilistic Model

The model proposed by Li is particularly advantageous in a number of ways; however, in terms of our interest in synthetic language, the model makes a number of assumptions which depart from known statistical linguistics. For example, the model assumes that for a word of length L, the total number of words possible is M^L given by (28). However, for a typical alphabet size, this vastly overestimates the number of words expected in a language, including many words which would not occur in known human languages.

As described in the previous section, the problem with the current ZML law is that the model is based on a simple estimate of the upper limit of possible words without consideration given to linguistic rules or other natural language principles beyond the initial power law. For example, in the English language, this might include orthographic spelling rules such as: (a) every word has at least one vowel; (b) "q" is almost always followed by "u"; (c) "s" never follows "x"; and (d) words never end in "v" or "j". These could potentially be considered as priors in a model, and there are other aspects of interaction in human communication and natural languages beyond linguistics which could be considered as statistical principles (For convenience we refer to these broadly as linguistic constraints and note that they may be related to verbal or written language.)to include in a model.

Another view of this problem is in terms of optimal coding, where the aim is typically to encode the most frequently used words in the most efficient way [67,68]. Despite ongoing interest in this area, there is strong evidence for Zipf's law of abbreviation which indicates that the highest ranking most frequent words tend to be shorter [38].

In contrast to a considerable body of work in deriving models to analyze and understand natural language, our interest is in deriving a probabilistic framework for constructing synthetic languages. Hence, in this section, we propose to consider a modified cZML law which is constrained to include linguistic principles.

As an example of linguistic features, we might consider the example of double-letter words. An extremely small number of words have double letters in comparison to the number of actual words possible. For five-letter English words, there are approximately 100 readily identifiable double letter words out of a possible 1.5×10^{18} words. Hence, we can introduce a new ZML model which introduces the constraint of not permitting words with adjacent double letters.

The derivation of the ZML with linguistic constraints is given in Appendix A.1. The effect of parametrization due to the constraints is indicated in Figure 3 where it can be readily seen that the effect is most significant for smaller values of M but diminishes quickly as M increases. Similarly, the effect of the constrained cZML model can be observed in

Figure 4 where the change in the symbolic probabilities can be observed for small values of M.

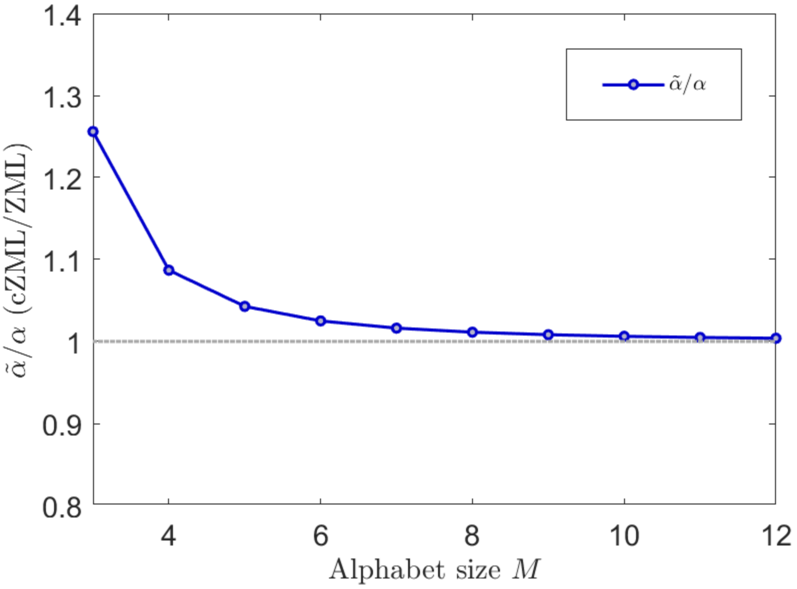

Figure 3. The effect of the proposed linguistic constraints on the modified ZML (cZML) vs. the usual ZML model are shown here by contrasting the relative parameters $\tilde{\alpha}/\alpha$ against M. It can be readily seen that the effect is most significant for smaller values of M but diminishes quickly as M increases, observed here for values up to $M = 12$.

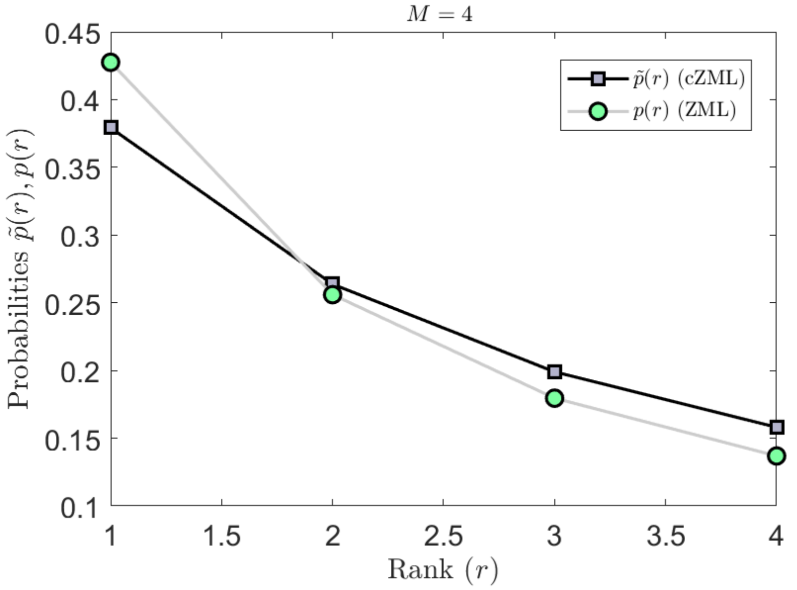

Figure 4. The effect of the proposed linguistic constraints on the modified cZML model are shown here by contrasting the ranked probabilities in each case. An example is shown here for $M = 4$.

This new cZML model introduces synthetic linguistic constraints based on having no adjacent repeating symbols. It is evident that other constraints can be considered to improve the accuracy of the model from a linguistics perspective, which we derive in the next section.

3.4. Constrained Linguistic Probabilistic Model II

This issue of the language space is well known in terms of language smoothing, where techniques such as the CN-gram (continuation n-gram [69]) have been proposed to reduce the space of possible words. Based on known human languages, the value of $\widetilde{N}_w(L;M)$ used by the constrained ZML model considered above is generally too large for a given value of M.

Hence, in this section, we derived a constrained cZML which introduces a reduced lexicon space through a CN-gram-style approach. The derivation of the ZML with linguistic constraints in this case is given in Appendix A.2.

The effect of the second form of the constrained ZML model can be observed in Figure 5 where the change in the symbolic probabilities can be observed for small values of M. While only a small probabilistic variation is observed, this corresponds to a significantly reduced maximum vocabulary size, which will lead to a more accurate estimation of entropy.

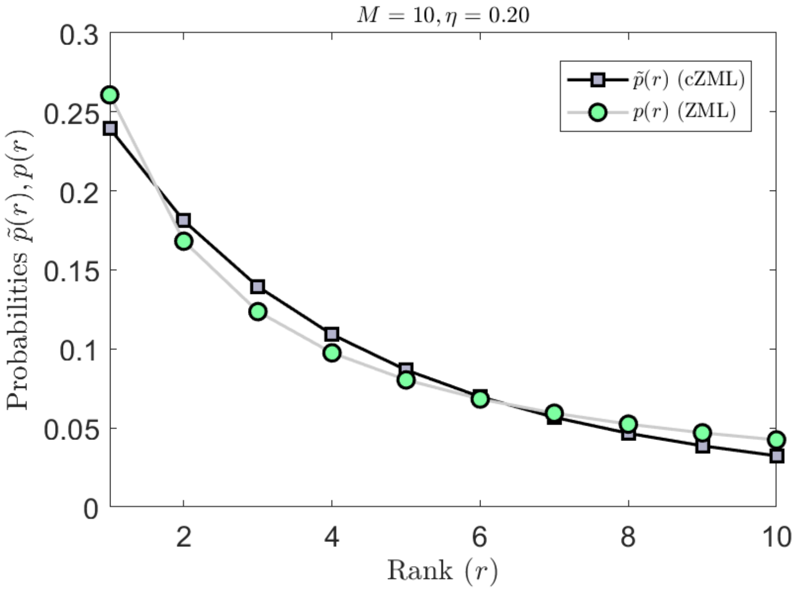

Figure 5. The second form of linguistically constrained cZML model has the effect of flattening the ranked probabilities. In the example shown here for $M = 10$, it can be observed that the mid-ranked probabilities are increased, while the highest and lowest ranked probabilities are decreased.

In the next section, we consider the performance of these newly proposed constrained ZML models within the efficient model-based entropy estimation algorithm described in Section 2.

4. Performance Results

4.1. Constrained Linguistic ZML Model for Natural Language

To test the performance of the proposed models, we applied them to English language data from the Google Web Trillion Word Corpus [70,71].

The proposed linguistically constrained cZML model approximates the higher ranked probabilities with better accuracy than the original ZML model and also enables better accuracy for the low-ranked probabilities (Figures 6 and 7).

We consider the full set of 676 two letter bigrams from the Google data set. The nonlinear behavior of the actual data is evident (Figure 7). The original ZML model shows linear behavior and does not approximate the low ranked probabilities very well. In contrast, the proposed linguistically constrained cZML model has the effect of flattening the ranked probabilities to give better high-ranked approximation, while also producing more accurate behavior for the low ranked probabilities.

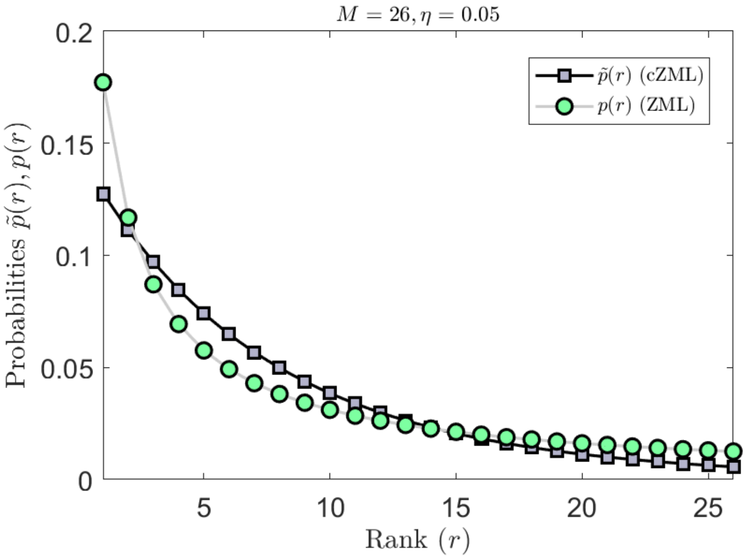

Figure 6. The second form of linguistically constrained cZML model has the effect of flattening the ranked probabilities. In the example shown here for $M = 26$, it can be observed that the mid-ranked probabilities are increased, while the highest and lowest ranked probabilities are decreased.

Note that since we utilize a model-based approach for estimating entropy, the choice of Zipfian model is significant to the outcome. Hence, because the proposed model more accurately approximates typical linguistic sequences, this can can be expected to lead to a more accurate entropy estimation process, though still within the limitations described above.

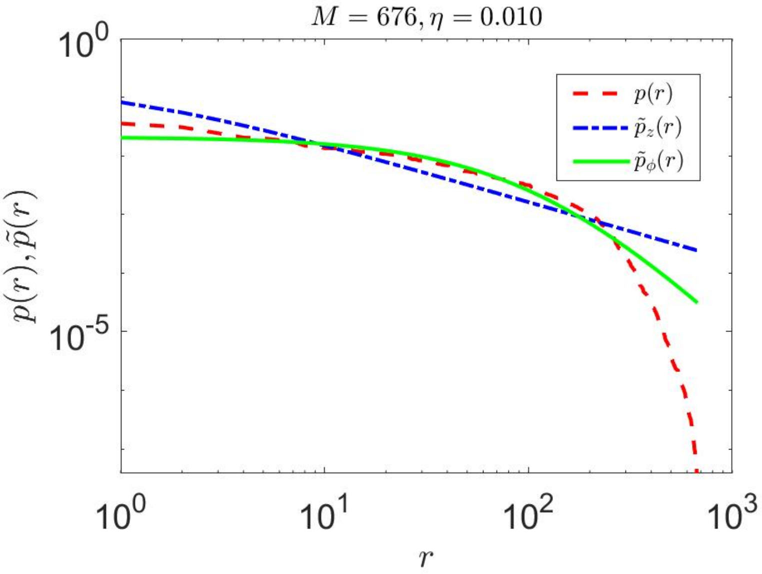

Figure 7. Performance of the constrained linguistic cZML model on actual English language data as compared to the unconstrained ZML model. In this case, we consider the full set of 676 two letter bigrams from the Google data set. The nonlinear behavior of the actual data is evidently (dashed curve) modeled with a higher degree of accuracy than the unconstrained model (dot dashed curve).

It is evident that, given the success of this approach, it is possible to consider numerous other constraints to improve the accuracy of the model from a linguistic perspective.

4.2. Entropy Rate Estimation

The entropy rate can be defined as the limit of joint entropy for an increasing number of symbols given by [67]

$$h_r(X) = \lim_{N\to\infty} \frac{1}{N} H_N(X) \tag{34}$$

$$= \lim_{N\to\infty} \frac{1}{N} H_N(X_1, \ldots, X_N) \tag{35}$$

where $X = X_1, \ldots, X_N$ can represent successive blocks of symbols. The task of entropy rate estimation is known to present a challenge due to the difficulty in obtaining a consistent estimate [72] and various number methods have been described. It is shown that, for the condition of stationarity, the entropy rate can be defined in terms of conditional entropy as [67]:

$$h_r(X) = \lim_{N\to\infty} H_N(X_N|X_1, \ldots, X_{N-1}) \tag{36}$$

In practice, for finite sequences of symbols, the condition of stationarity may not hold and therefore estimating the entropy rate using (35) may result in a different value than with (36).

The value of entropy rate was estimated by Shannon using an experimental approach and found to be about one bit-per-character (bpc) [5]. Cover estimated the entropy rate for the novel *Jefferson the Virginian*, by Dumas Malone to be 1.25 bpc [73]. A word trigram method which used the cross-entropy between this model and a balanced sample of English text trained on a language model of 583 million symbols was applied to Form C of the Brown corpus which yielded an upper bound entropy rate estimate of 1.75 bpc [74]. A number of unigram entropy estimation methods using a stabilization criterion and a linear entropy to entropy rate conversion model were considered in terms of a large scale study across three parallel corpora, encompassing approximately 450M words in 1259 languages, leading to estimates of the entropy rate of 6 bits per word in [75]. An estimation method using the limit of successive backward differences in n-gram entropies was proposed in [76]. Compression algorithms have also been used as a basis for entropy rate estimation [77,78].

A method using an exponential extrapolation function was proposed in [79] to provide an estimate of entropy rate across multiple languages and 20 corpora provided results tending towards infinity. Interestingly, this result indicated that the entropy rates of human languages are positive but approximately 20% smaller than without extrapolation, which appears to be in agreement with the results obtained for entropy rate estimation using the algorithm proposed here.

Some issues are evident in the experimental studies because of the use of different entropy rate estimation algorithms, different corpora, the treatment of how n-grams are evaluated, for example whether only actually occurring n-grams within words are used or not (see for example the contrasting discussions between [3,74,75]), the inclusion of only alphabetic characters or whether to include punctuation, and various other factors.

Here, we estimate the entropy rate using the proposed cZML model-based entropy estimator applied to the Brown corpus which consists of approximately 5.5 million characters [80]. While a comprehensive comparison of the various entropy estimation algorithms is beyond the scope of this paper, the results of the proposed algorithm are compared with the plug-in entropy estimation approach. Prefiltering was performed to remove all non-alphanumeric characters except for spaces, and n-grams which are not part of any word were excluded. It is well recognized that a smaller data set presents a challenge for entropy estimation especially with increasing word lengths [3,75] and so it is of interest to observe the relative performance of the proposed algorithm.

The conditional entropy rate estimation approach of (36) was adopted in each case. For the entire Brown corpus, the entropy rate was estimated as 1.29 bpc, whereas using plug-in method, the result was 2.04 bpc and outside the upper bound indicated in [74].

The results from the proposed algorithm compare well with the result of 1.25 bpc for a large collection of English corpora in [3]; however, it is evident that the results from the plug-in method are significantly different, giving less confidence in their reliability. The advantage of requiring fewer samples for the proposed entropy estimation algorithm is also apparent in this case of entropy rate estimation.

4.3. Convergence of Constrained cZML Entropy Estimation Algorithm

The convergence characteristics of the proposed algorithm can be thereby obtained by generating a random sequence of symbols according to $\{p_j(x)\}$ and the estimated entropy is developed as a function of the sample size N_s.

This approach is applied to an example case where $M = 30$ with the results shown in Figure 8. The performance of the proposed algorithm is compared with a conventional plug-in estimator on a defined entropy estimation task over a large range of samples sizes up to $N_s = 10^6$ symbolic samples. The mean entropy estimate is obtained by averaging over $N_v = 75$ trials, where it can be observed that the new algorithm converges very rapidly, requiring significantly fewer samples to converge compared to a conventional plug-in estimator approach [14].

Note that since both the proposed algorithm and the conventional plug-in entropy estimation algorithm rely on a maximum likelihood method, the computational burden of each is $O(N)$. However, the key advantage of the proposed model-based algorithm is that it requires substantially fewer samples than the conventional plug-in entropy estimator.

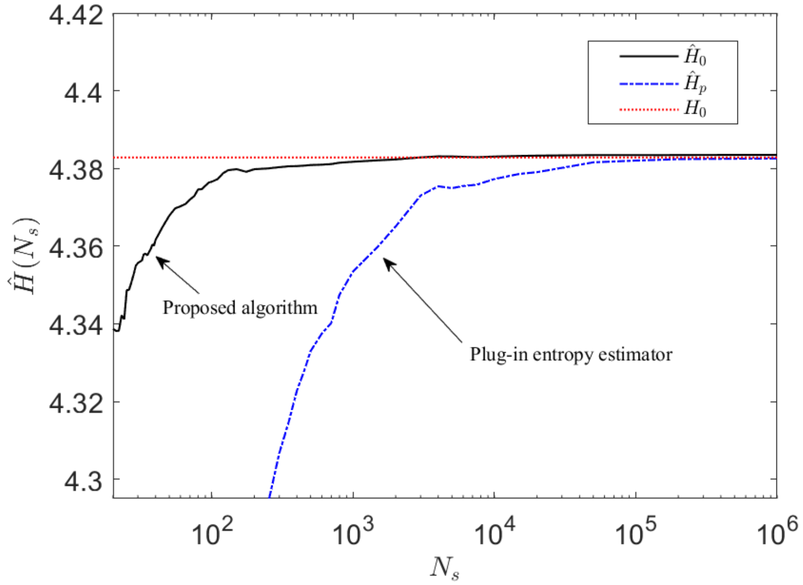

Figure 8. The convergence performance of the proposed entropy estimation algorithm is compared to a regular plug-in estimator for $M = 30$ as a function of the sample size, up to $N_s = 10^6$ samples averaged across $N_v = 75$ trials. It can be noted that the estimate $\widehat{H}_0(N_s)$ from the proposed algorithm converges rapidly very closely towards the true entropy value (within three decimal places). In contrast, the conventional plug-in estimator $\widehat{H}_p(N_s)$ converges much more slowly towards a biased estimate.

It is well known that conventional plug-in estimators will give biased estimates. For the proposed algorithm, any such bias will result from the accuracy with which $D_r(M)$ can be estimated. It can be noted that as with conventional probability estimates, provided the system is stationary, it is possible to improve the estimate of $D_r(M)$ by increasing the number of samples.

5. Conclusions

Entropy estimators which go beyond naive maximum likelihood methods are of considerable interest, particularly in terms of overcoming limitations due to data. The approach of coincidence counting is recognized as a potentially powerful approach for model-based estimators. Here, we show that an efficient coincidence counting estimator can be derived using a Zipf–Mandelbrot–Li law which provides a significant reduction in the data required.

Interestingly, while Zipfian laws are ubiquitous in fields such as natural language, surprisingly, it appears that these models have evidently been developed essentially without particular regard for the possible inclusion of linguistic constraints.

In this paper, by introducing some simple linguistic constraints, we extended the regular rank-based Zipf–Mandelbrot–Li model to one which provides more realistic assumptions and makes it more suitable for a broad range of probabilistic language models which rely on an analytical Zipfian framework. Such models can be applied to the human language and provide a necessary foundation for developing synthetic language models.

Conceptually, the idea of model-based entropy estimators seems like it may potentially sacrifice accuracy; however, our results show that for natural systems where the symbolic events follow an approximate Zipfian distribution and using limited data, the performance is better than that obtained by a naive entropy estimator. We derived results which indicate that the expected improvement in convergence and demonstrated the efficacy of the proposed model on the entropy rate estimation and two-letter bigram entropy estimation where it was shown to produce more accurate behavior for both low-ranked and high-ranked symbolic probabilities.

The proposed constrained linguistic Zipf–Mandelbrot–Li model appears to be the first time this approach has been adopted. In future work, it would be of interest to further extend this concept by introducing more sophisticated linguistic constraints and to explore applications where limited data are available.

Author Contributions: Data curation, A.D.B.; Formal analysis, A.D.B.; Funding acquisition, A.D.B. and J.W.; Investigation, A.D.B. and J.W.; Methodology, A.D.B. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The University of Queensland and Trusted Autonomous Systems Defence Cooperative Research Centre, Grant Number 2019002828.

Data Availability Statement: Data used for Section 4.1, derived from the Google Web Trillon Word Corpus is available from https://norvig.com/ngrams (accessed on 18 August 2021). The Brown corpus used in Section 4.2 is available as part of the NLTK from https://www.nltk.org/book/ch02.html (accessed on 18 August 2021) .

Acknowledgments: The authors gratefully acknowledge funding support from the University of Queensland and from the Australian Government through the Defence Cooperative Research Centre for Trusted Autonomous Systems. The DCRC-TAS receives funding support from the Queensland Government.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Derivation of Constrained Linguistic Zipf–Mandelbrot–Li Models

Appendix A.1. cZML Model I

The Zipf–Mandelbrot–Li law with linguistic constraints of the first type is derived as follows. For a word of length L, which is constrained to have no repeating adjacent symbols, the total number of words possible is given by

$$\begin{aligned}\widetilde{N}_w &= M \prod_{k=1M}^{L} (M_k - 1), \quad k = 2, \ldots, L \\ &= M(M-1)^{L-1} \end{aligned} \tag{A1}$$

and the frequency of occurrence for a word of length L using the constrained probabilistic model is:

$$\widetilde{p}_i(L) = \frac{\widetilde{\gamma}}{(M+1)M^{L+1}} \qquad i = 1, \ldots, \widetilde{N}_w(L) \tag{A2}$$

where $\widetilde{\gamma}$ is a normalization constant. It follows that $\widetilde{\gamma}$ can be determined from the word frequency normalization condition via the summation of all probabilities of such words [30], hence:

$$\sum_{L=1}^{\infty} \widetilde{N}_w(L)\widetilde{p}_i(L) = 1 \tag{A3}$$

$$= \sum_{L=1}^{\infty} \frac{\widetilde{\gamma}(M-1)^{L-1}}{(M+1)M^L} \tag{A4}$$

$$= \frac{\widetilde{\gamma}}{(M+1)} \tag{A5}$$

and hence the normalization constant can be found as

$$\widetilde{\gamma} = M + 1 \tag{A6}$$

From (A2), the probability for any possible constrained word is:

$$\widetilde{p}_i(L) = \frac{1}{M^{L+1}} \qquad i = 1, \ldots, \widetilde{N}_w(L) \tag{A7}$$

and hence the frequency of occurrence of all words of length L and with the constraint that there are no adjacent repeating symbols is given by an exponential function of L as

$$\widetilde{p}(L) = M(M-1)^{L-1}\widetilde{p}_i(L) \tag{A8}$$

$$= \frac{(M-1)^{L-1}}{M^L} \tag{A9}$$

Now it follows that the rank can be considered either in terms of the direct probability or equivalently, as the incremental change in probability as we change the word length, i.e., from $L-1$ to L. Consider the rank of probabilities in the exponentially decreasing distribution $\{p_i(M,L)\}$, defined as

$$r(i; \{p_i(M,L)\}) = \arg_i(\{p_i(M,L)\}) \tag{A10}$$

then from (A2) it follows that:

$$p_i(L) < p_i(L+1) \tag{A11}$$

and hence:

$$r(L) > r(L+1) \tag{A12}$$

where it is evident that the rank $r(L)$ is proportional to an inverse function g of the corresponding probabilities, such that:

$$\begin{aligned} r(L) &= g(M,L) \\ &\leq M^L \end{aligned} \tag{A13}$$

Hence, since the rank is effectively determined by the inverse of the probabilities, and the probabilities are found as the inverse of the number of occurrences, and accordingly, we can calculate the total number of cumulative occurrences for all words up to length L, which in turn provides a measure of the ranking. Therefore, we have:

$$\sum_{k=1w}^{L-1} \widetilde{N}_w(k) < r(L) \leq \sum_{k=1}^{L} \widetilde{N}_w(k) \tag{A14}$$

where:

$$\sum_{k=1}^{L-1} \widetilde{N}_w(k) = \sum_{k=1}^{L-1} M(M-1)^{k-1} \tag{A15}$$

$$= \frac{M\Big((M-1)^{L-1}-1\Big)}{M-2} \tag{A16}$$

and:

$$\sum_{k=1}^{L} \widetilde{N}_w(k) = \sum_{k=1}^{L} M(M-1)^{k-1} \tag{A17}$$

$$= \frac{M\Big((M-1)^{L}-1\Big)}{M-2} \tag{A18}$$

and hence:

$$\frac{M\Big((M-1)^{L-1}-1\Big)}{M-2} < r(L) \leq \frac{M\Big((M-1)^{L}-1\Big)}{M-2} \tag{A19}$$

which represents the exponential transformation from the constrained word's length to the word's rank under the given constraint:

Rearranging (A19) and taking logs, for the lower bound, we have:

$$\frac{(M-2)r(L)}{M} > (M-1)^{L-1} - 1 \tag{A20}$$

$$L - 1 < \log_{M-1}\left(\frac{(M-2)r(L)}{M} + 1\right) \tag{A21}$$

and for the upper bound, we have:

$$\frac{(M-2)r(L)}{M} \leq (M-1)^{L} - 1 \tag{A22}$$

$$L \geq \log_{M-1}\left(\frac{(M-2)r(L)}{M} + 1\right) \tag{A23}$$

Following a similar approach to [30], raising $\frac{1}{M}$ to the power of the terms in (A21)–(A23) gives:

$$\frac{1}{M^{L-1}} > \left(\frac{1}{M}\right)^{\varsigma} \geq \left(\frac{1}{M^{L}}\right) \tag{A24}$$

where:

$$\varsigma = \log_{M-1}\left(\frac{(M-2)r(L)}{M} + 1\right) \tag{A25}$$

Using the identity

$$\left(\frac{1}{a}\right)^{\log b} = \left(\frac{1}{b}\right)^{\log a} \tag{A26}$$

then, it follows using a change of base:

$$\left(\frac{1}{M}\right)^{\varsigma} = \left(\frac{1}{(\frac{M-2}{M})r(L)+1}\right)^{\log_{M-1}(M)} \tag{A27}$$

$$= \left(\frac{1}{(\frac{M-2}{M})r(L)+1}\right)^{\frac{\log(M)}{\log(M-1)}} \tag{A28}$$

noting that:

$$\frac{1}{MM^{L-1}} = \frac{1}{M^L} = \widetilde{p}_i(L-1) \tag{A29}$$

$$\frac{1}{MM^{L}} = \frac{1}{M^{L+1}} = \widetilde{p}_i(L) \tag{A30}$$

then (A24) can be multiplied by $1/M$ to give probabilistic bounds, and hence we have:

$$\frac{1}{M}\left(\frac{1}{(\frac{M-2}{M})r(L)+1}\right)^{\widetilde{\alpha}} = \frac{\frac{1}{M}(\frac{M}{M-2})^{\widetilde{\alpha}}}{\left(r(L)+\frac{M}{M-2}\right)^{\widetilde{\alpha}}} \tag{A31}$$

$$= \frac{\frac{M^{\widetilde{\alpha}-1}}{(M-2)^{\widetilde{\alpha}}}}{\left(r(L)+\frac{M}{M-2}\right)^{\widetilde{\alpha}}} \tag{A32}$$

$$= \frac{\widetilde{\gamma}}{\left(r(L)+\widetilde{\beta}\right)^{\widetilde{\alpha}}} \tag{A33}$$

where:

$$\widetilde{\alpha} = \frac{\log(M)}{\log(M-1)} \tag{A34}$$

$$\widetilde{\beta} = \frac{M}{M-2} \tag{A35}$$

$$\widetilde{\gamma} = \frac{M^{\widetilde{\alpha}-1}}{(M-2)^{\widetilde{\alpha}}} \tag{A36}$$

which leads to:

$$\widetilde{p}_i(L-1) < \frac{\widetilde{\gamma}}{\left(r(L)+\widetilde{\beta}\right)^{\widetilde{\alpha}}} \leq \widetilde{p}_i(L) \tag{A37}$$

which can be considered as a new form of Zipf–Mandelbrot–Li law which includes the specified linguistic constraints such that the constants originally computed according to (12)–(16), are now computed according to (A34)–(A36).

Note that as the alphabet size M increases, the parameter given by (A34) converges to that of (12) in the original formulation:

$$\lim_{M\to\infty}\frac{\widetilde{\alpha}}{\alpha} = 1 \tag{A38}$$

A formulation of the model using a maximum word length $L_{\max}$ can now be introduced. It follows that a new value for $\widetilde{\gamma}$ can be determined from the word frequency normalization condition via the summation of all probabilities as before, hence:

$$\sum_{L=1}^{L_{\max}} \widetilde{N}_w(L)\widetilde{p}_i(L) = 1 \tag{A39}$$

$$= \sum_{L=1}^{L_{\max}} \frac{\widetilde{\gamma}(M-1)^{L-1}}{(M+1)M^L} \tag{A40}$$

which can be shown to reduce to:

$$\frac{\widetilde{\gamma}}{M+1} - \frac{\widetilde{\gamma}M(\frac{M-1}{M})^{(L_{\max}+1)}}{(M+1)(M-1)} = 1 \tag{A41}$$

$$= \frac{\widetilde{\gamma}\Big(1-(\frac{M-1}{M})^{L\max}\Big)}{M+1} \tag{A42}$$

and hence:

$$\widetilde{\gamma} = \frac{M+1}{1-(\frac{M-1}{M})^{L_{\max}}} \tag{A43}$$

Appendix A.2. cZML Model II

The Zipf–Mandelbrot–Li law with linguistic constraints of the second type is derived as follows. For a word of length L, the total number of words possible is given by

$$\widetilde{N}_w = M\prod_{k=1}^{L-1}\eta(M_k-1), \quad k = 1,\ldots,L$$

$$= \eta^{L-1}M(M-1)^{L-1} \tag{A44}$$

and hence the frequency of occurrence for a word of length L using the constrained probabilistic model is

$$\widetilde{p}_i(L) = \frac{\widetilde{\gamma}}{(M+1)(\eta M)^{L+1}} \qquad i = 1,\ldots,\widetilde{N}_w(L) \tag{A45}$$

where: $\widetilde{\gamma}$ is a normalization constant. It follows that $\widetilde{\gamma}$ can be determined from the word frequency normalization condition via the summation of all probabilities of such words [30], hence:

$$\sum_{L=1}^{\infty} \widetilde{N}_w(L)\widetilde{p}_i(L) = 1 \tag{A46}$$

$$= \sum_{L=1}^{\infty} \frac{\widetilde{\gamma}\eta^{L-1}M(M-1)^{L-1}}{(M+1)(\eta M)^{L+1}} \tag{A47}$$

$$= \sum_{L=1}^{\infty} \frac{\widetilde{\gamma}\eta^{-2}M^{-L}(M-1)^{L-1}}{(M+1)} \tag{A48}$$

$$= \frac{\widetilde{\gamma}}{\eta^2(M+1)} \tag{A49}$$

and hence the normalization constant can be found as

$$\widetilde{\gamma} = \eta^2(M+1) \tag{A50}$$

From (A45), the probability for any possible constrained word is

$$\widetilde{p}_i(L) = \frac{1}{\eta^{L-1}M^{L+1}} \qquad i = 1, \dots, \widetilde{N}_w(L) \tag{A51}$$

and hence the frequency of the occurrence of all words of length L and with the constraint that there are no adjacent repeating symbols is given by an exponential function of L as

$$\begin{aligned}\widetilde{p}(L) &= \eta^{L-1}M(M-1)^{L-1}\widetilde{p}_i(L) && \text{(A52)}\\ &= \frac{(M-1)^{L-1}}{M^L} && \text{(A53)}\end{aligned}$$

Now it follows that the rank can be considered either in terms of direct probability or equivalently, as the incremental change in probability as we change the word length, i.e., from $L-1$ to L. Consider the rank of probabilities in the exponentially decreasing distribution $\{p_i(M,L)\}$, defined as

$$r(i; \{p_i(M,L)\}) = \arg_i(\{p_i(M,L)\}) \tag{A54}$$

then, from (A45), it follows that:

$$p_i(L) < p_i(L+1) \tag{A55}$$

and hence:

$$r(L) > r(L+1) \tag{A56}$$

where it is evident that the rank $r(L)$ is proportional to an inverse function g of the corresponding probabilities such that:

$$\begin{aligned}r(L) &= g(M,L)\\ &\leq M^L && \text{(A57)}\end{aligned}$$

Hence, since the rank is effectively determined by the inverse of the probabilities, and the probabilities are found as the inverse of the number of occurrences, accordingly, we can calculate the total number of cumulative occurrences for all words up to length L, which in turn provides a measure of the ranking. Therefore, we have:

$$\sum_{k=1w}^{L-1} \widetilde{N}_w(k) < r(L) \leq \sum_{k=1}^{L} \widetilde{N}_w(k) \tag{A58}$$

where:

$$\begin{aligned}\sum_{k=1}^{L-1} \widetilde{N}_w(k) &= \sum_{k=1}^{L-1} \eta^{k-1}M(M-1)^{k-1} && \text{(A59)}\\ &= \frac{M\left(\eta^L(M-1)^L - \eta(M-1)\right)}{\eta(M-1)(\eta(M-1)-1)} && \text{(A60)}\\ &= \frac{M\left(\eta^{L-1}(M-1)^{L-1} - 1\right)}{\eta(M-1)-1} && \text{(A61)}\end{aligned}$$

and:

$$\sum_{k=1}^{L}\widetilde{N}_w(k) = \sum_{k=1}^{L}\eta^{k-1}M(M-1)^{k-1} \tag{A62}$$

$$= \frac{M(\eta^L(M-1)^L-1)}{\eta(M-1)-1} \tag{A63}$$

and hence:

$$\frac{M\Big(\eta^{L-1}(M-1)^{L-1}-1\Big)}{\eta(M-1)-1} < r(L) \leq \frac{M(\eta^L(M-1)^L-1)}{\eta(M-1)-1} \tag{A64}$$

which represents the exponential transformation from the constrained word's length to the word's rank under the given constraint. Rearranging (A64) and taking logs, we have:

$$\frac{(\eta(M-1)-1)r(L)}{M} > \eta^{L-1}(M-1)^{L-1}-1 \tag{A65}$$

$$L-1 < \log_{\eta(M-1)}\left(\frac{(\eta(M-1)-1)r(L)}{M}+1\right) \tag{A66}$$

and for the upper bound, we have:

$$\frac{(\eta(M-1)-1)r(L)}{M} \leq (M-1)^L-1 \tag{A67}$$

$$L \geq \log_{\eta(M-1)}\left(\frac{(\eta(M-1)-1)r(L)}{M}+1\right) \tag{A68}$$

Following a similar approach to [30], raising $\frac{1}{M}$ to the power of the bound terms in (A65)–(A68) gives:

$$\frac{1}{M^{L-1}} > (\frac{1}{M})^{\varsigma} \geq (\left(\frac{1}{M^L}\right)) \tag{A69}$$

where:

$$\varsigma = \log_{\eta(M-1)}\left(\frac{(\eta(M-1)-1)r(L)}{M}+1\right) \tag{A70}$$

Using the identity:

$$\left(\frac{1}{a}\right)^{\log b} = \left(\frac{1}{b}\right)^{\log a} \tag{A71}$$

then it follows using a change of base:

$$\left(\frac{1}{M}\right)^{\varsigma} = \left(\frac{1}{(\frac{\eta(M-1)-1}{M})r(L)+1}\right)^{\log_{\eta(M-1)}(M)} \tag{A72}$$

$$= \left(\frac{1}{(\frac{\eta(M-1)-1}{M})r(L)+1}\right)^{\frac{\log(M)}{\log\eta(M-1)}} \tag{A73}$$

Noting that:

$$\frac{1}{MM^{L-1}} = \frac{1}{M^L} = \widetilde{p}_i(L-1) \tag{A74}$$

$$\frac{1}{MM^L} = \frac{1}{M^{L+1}} = \widetilde{p}_i(L) \tag{A75}$$

then (A69) can be multiplied by $1/M$ to give probabilistic bounds, and hence we have:

$$\frac{1}{M}\left(\frac{1}{(\frac{\eta(M-1)-1}{M})r(L)+1}\right)^{\widetilde{\alpha}} = \frac{\frac{1}{M}(\frac{M}{\eta(M-1)-1})^{\widetilde{\alpha}}}{\left(r(L)+\frac{M}{\eta(M-1)-1}\right)^{\widetilde{\alpha}}} \tag{A76}$$

$$= \frac{\frac{M^{\widetilde{\alpha}-1}}{(\eta(M-1)-1)^{\widetilde{\alpha}}}}{\left(r(L)+\frac{M}{\eta(M-1)-1}\right)^{\widetilde{\alpha}}} \tag{A77}$$

$$= \frac{\widetilde{\gamma}}{\left(r(L)+\widetilde{\beta}\right)^{\widetilde{\alpha}}} \tag{A78}$$

where:

$$\widetilde{\alpha} = \frac{\log(M)}{\log \eta(M-1)} \tag{A79}$$

$$\widetilde{\beta} = \frac{M}{\eta(M-1)-1} \tag{A80}$$

$$\widetilde{\gamma} = \frac{M^{\widetilde{\alpha}-1}}{(\eta(M-1)-1)^{\widetilde{\alpha}}} \tag{A81}$$

which leads to:

$$\widetilde{p}_i(L-1) < \frac{\widetilde{\gamma}}{\left(r(L)+\widetilde{\beta}\right)^{\widetilde{\alpha}}} \leq \widetilde{p}_i(L) \tag{A82}$$

which can be considered as a further linguistically constrained form of the Zipf–Mandelbrot–Li law, which includes the specified linguistic constraints such that the constants originally computed according to (12)–(16) are now computed according to (A79)–(A81).

Note that as before, as the alphabet size M increases, the parameter given by (A79) converges to that of (12) in the original formulation:

$$\lim_{M\to\infty} \frac{\widetilde{\alpha}}{\alpha} = 1 \tag{A83}$$

and the effect of the parametrization due to the constraints is similar to the first case.

Following the same approach as before, a formulation of the model using a maximum word length $L_{\max}$ can now be introduced. It follows that a new value for $\widetilde{\gamma}$ can be determined from the word frequency normalization condition via the summation of all probabilities as before, hence:

$$\sum_{L=1}^{L_{\max}} \widetilde{N}_w(L)\widetilde{p}_i(L) = 1$$

$$= \sum_{L=1}^{L_{\max}} \frac{\widetilde{\gamma}(M-1)^{L-1}}{(M+1)M^L} \tag{A84}$$

$$= \sum_{L=1}^{L_{\max}} \frac{\widetilde{\gamma}\eta^{-2}M^{-L}(M-1)^{L-1}}{(M+1)} \tag{A85}$$

which can be shown to reduce to:

$$\frac{\widetilde{\gamma}}{M+1} - \frac{\widetilde{\gamma}M(\frac{M-1}{M})^{(L_{\max}+1)}}{(M+1)(M-1)} = 1 \tag{A86}$$

$$\frac{\widetilde{\gamma}\left(1-(\frac{M-1}{M})^{L\max}\right)}{\eta^2(M+1)} = 1 \tag{A87}$$

and hence:

$$\widetilde{\gamma} = \frac{\eta^2(M+1)}{1-(\frac{M-1}{M})^{L_{\max}}} \tag{A88}$$

This ZML model introduces further synthetic linguistic constraints based on having no adjacent repeating symbols.

References

1. Shannon, C.E. A Mathematical Theory of Communication (Parts I and II). *Bell Syst. Tech. J.* **1948**, *XXVII*, 379–423. [CrossRef] [CrossRef]
2. Shannon, C.E. A Mathematical Theory of Communication (Part III). *Bell Syst. Tech. J.* **1948**, *XXVII*, 623–656. [CrossRef] [CrossRef]
3. Schürmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos* **1996**, *6*, 414–427. [CrossRef] [CrossRef] [PubMed]
4. Jelinek, F.; Mercer, R.L.; Bahl, L.R.; Baker, J.K. Perplexity—A measure of the difficulty of speech recognition tasks. *J. Acoust. Soc. Am.* **1977**, *62*, S63. [CrossRef] [CrossRef]
5. Shannon, C.E. Prediction and Entropy of Printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64. [CrossRef] [CrossRef]
6. Ziv, J.; Lempel, A. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory* **1978**, *24*, 530–536. [CrossRef] [CrossRef]
7. Amigó, J.; Szczepanski, J.; Wajnryb, E.; Sanchez-Vives, M. Estimating the Entropy Rate of Spike Trains via Lempel-Ziv Complexity. *Neural Comput.* **2004**, *16*, 717–736. [CrossRef] [CrossRef] [PubMed]
8. Porta, A.; Guzzetti, S.; Montano, N.; Furlan, R.; Pagani, M.; Malliani, A.; Cerutti, S. Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series. *IEEE Trans. Biomed. Eng.* **2001**, *48*, 1282–1291. [CrossRef] [CrossRef]
9. Wang, W.; Wang, Y.; Huang, Q.; Gao, W. Measuring visual saliency by Site Entropy Rate. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2368–2375.
10. Kershenbaum, A. Entropy rate as a measure of animal vocal complexity. *Bioacoustics* **2014**, *23*, 195–208. [CrossRef] [CrossRef]
11. Loewenstern, D.; Yianilos, P.N. Significantly Lower Entropy Estimates for Natural DNA Sequences. *J. Comput. Biol.* **1999**, *6*, 125–142. [CrossRef] [CrossRef]
12. Vegetabile, B.G.; Stout-Oswald, S.A.; Davis, E.P.; Baram, T.Z.; Stern, H.S. Estimating the Entropy Rate of Finite Markov Chains With Application to Behavior Studies. *J. Educ. Behav. Stat.* **2019**, *44*, 282–308. [CrossRef] [CrossRef]
13. Braverman, M.; Chen, X.; Kakade, S.; Narasimhan, K.; Zhang, C.; Zhang, Y. Calibration, Entropy Rates, and Memory in Language Models. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; Hal, D., III, Singh, A., Eds.; Proceedings of Machine Learning Research; ML Research Press: Vienna, Austria, 2020; Volume 119, pp. 1089–1099.
14. Back, A.D.; Angus, D.; Wiles, J. Determining the Number of Samples Required to Estimate Entropy in Natural Sequences. *IEEE Trans. Inf. Theory* **2019**, *65*, 4345–4352. [CrossRef] [CrossRef]
15. Lesne, A.; Blanc, J.L.; Pezard, L. Entropy estimation of very short symbolic sequences. *Phys. Rev. E* **2009**, *79*, 046208. [CrossRef] [PubMed] [CrossRef]
16. Hausser, J.; Strimmer, K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.* **2009**, *10*, 1469–1484.
17. Wolpert, D.H.; Wolf, D.R. Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E* **1995**, *52*, 6841–6854. [CrossRef] [PubMed] [CrossRef]
18. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and Inference, Revisited. In *Advances in Neural Information Processing Systems 14*; Dietterich, T.G., Becker, S., Ghahramani, Z., Eds.; MIT Press: Cambridge, MA, USA, 2002; pp. 471–478.
19. Montalvão, J.; Silva, D.; Attux, R. Simple entropy estimator for small datasets. *Electron. Lett.* **2012**, *48*, 1059–1061. [CrossRef] [CrossRef]
20. Bonachela, J.A.; Hinrichsen, H.; Muñoz, M.A. Entropy Estimates of Small Data Sets. *J. Phys. A Math. Theor.* **2008**, *41*, 1–9. [CrossRef] [CrossRef]
21. Paavola, M. An Efficient Entropy Estimation Approach. Ph.D. Thesis, University of Oulu, Oulu, Finland, 2011.
22. Gerlach, M.; Font-Clos, F.; Altmann, E.G. Similarity of Symbol Frequency Distributions with Heavy Tails. *Phys. Rev. X* **2016**, *6*, 021009. [CrossRef] [CrossRef]

23. Kugiumtzis, D. Partial Transfer Entropy on Rank Vectors. *Eur. Phys. J. Spec. Top.* **2013**, *222*, 401–420. [CrossRef] [CrossRef]
24. Paninski, L. Estimation of Entropy and Mutual Information. *Neural Comput.* **2003**, *15*, 1191–1253. [CrossRef] [CrossRef]
25. Kolchinsky, A.; Tracey, B.D. Estimating Mixture Entropy with Pairwise Distances. *Entropy* **2017**, *19*, 361. [CrossRef] [CrossRef]
26. Safaai, H.; Onken, A.; Harvey, C.D.; Panzeri, S. Information estimation using nonparametric copulas. *Phys. Rev. E* **2018**, *98*, 053302. [CrossRef] [PubMed] [CrossRef]
27. Hernández, D.G.; Samengo, I. Estimating the Mutual Information between Two Discrete, Asymmetric Variables with Limited Samples. *Entropy* **2019**, *21*, 623. [CrossRef] [CrossRef] [PubMed]
28. Ma, S. Calculation of Entropy from Data of Motion. *J. Stat. Phys.* **1981**, *26*, 221–240. [CrossRef] [CrossRef]
29. Montalvão, J.; Attux, R.; Silva, D. A pragmatic entropy and differential entropy estimator for small datasets. *J. Commun. Inf. Syst.* **2014**, *29*. [CrossRef] [CrossRef]
30. Li, W. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Trans. Inf. Theory* **1992**, *38*, 1842–1845. [CrossRef] [CrossRef]
31. Limpert, E.; Stahel, W.A.; Abbt, M. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience* **2001**, *51*, 341–352. [CrossRef] [CrossRef]
32. Giesbrecht, F.; Kempthorne, O. Maximum Likelihood Estimation in the Three-Parameter Lognormal Distribution. *J. R. Stat. Soc. Ser. B (Methodol.)* **1976**, *38*, 257–264. [CrossRef] [CrossRef]
33. Wang, S.; Gui, W. Corrected Maximum Likelihood Estimations of the Lognormal Distribution Parameters. *Symmetry* **2020**, *12*, 968. [CrossRef] [CrossRef]
34. Li1, B.; Yashchin, E.; Christiansen, C.; Gill, J.; Filippi, R.; Sullivan, T. Application of Three-Parameter Lognormal Distribution in EM Data Analysis. In *Mathematics IBM Research Report RC23680 (W0507-213)*; IBM Systems and Technology Group: Essex Junction, VT, USA, 2005.
35. Dvoretzky, A.; Kiefer, J.; Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **1956**, *27*, 642–669. [CrossRef] [CrossRef]
36. Zipf, G. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*; Houghton Mifflin: Cambridge, MA, USA, 1935.
37. Piantadosi, S.T. Zipf's word frequency law in natural language: A critical review and future directions. *Psychon. Bull. Rev.* **2014**, *21*, 1112–1130. [CrossRef] [CrossRef]
38. Bentz, C.; Ferrer-i-Cancho, R. Zipf's law of abbreviation as a language universal. In Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics, Lorentz Center, Leiden, 26–30 October 2015; Bentz, C., Jäger, G., Yanovich, I., Eds.; University of Tübingen: Tübingen, Germany, 2015.
39. Mantegna, R.N.; Buldyrev, S.V.; Goldberger, A.L.; Havlin, S.; Peng, C.K.; Simons, M.; Stanley, H.E. Linguistic Features of Noncoding DNA Sequences. *Phys. Rev. Lett.* **1994**, *73*, 3169–3172. [CrossRef] [CrossRef]
40. Zipf, G.; Thiele, L. *Human Behavior and the Principle of Least Effort*; Addison Wesley: Oxford, UK, 1949.
41. Miller, G.A. Some effects of intermittent silence. *Am. J. Psychol.* **1957**, *70*, 311–314. [CrossRef] [CrossRef] [PubMed]
42. Howes, D. Zipf's Law and Miller's Random-Monkey Model. *Am. J. Psychol.* **1968**, *81*, 269–272. [CrossRef] [CrossRef]
43. Conrad, B.; Mitzenmacher, M. Power laws for monkeys typing randomly: The case of unequal probabilities. *IEEE Trans. Inf. Theory* **2004**, *50*, 1403–1414. [CrossRef] [CrossRef]
44. Perline, R.; Perline, R. Two Universality Properties Associated with the Monkey Model of Zipf's Law. *Entropy* **2016**, *18*, 89. [CrossRef] [CrossRef]
45. Piantadosi, S.T.; Tily, H.; Gibson, E. Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 3526–3529. [CrossRef] [PubMed] [CrossRef] [PubMed]
46. Ferrer-i-Cancho, R.; Solé, R.V. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 788–791. [CrossRef] [PubMed] [CrossRef]
47. Gibson, E.; Futrell, R.; Piantadosi, S.P.; Dautriche, I.; Mahowald, K.; Bergen, L.; Levy, R. How Efficiency Shapes Human Language. *Trends Cogn. Sci.* **2019**, *23*, 389–407. [CrossRef] [PubMed] [CrossRef]
48. Steinert-Threlkeld, S.; Szymanik, J. Ease of learning explains semantic universals. *Cognition* **2020**, *195*, 104076. [CrossRef] [CrossRef]
49. Li, W. Zipf's Law Everywhere. *Glottometrics* **2002**, *5*, 14–21.
50. Corral, Á.; Boleda, G.; Ferrer-i-Cancho, R. Zipf's Law for Word Frequencies: Word Forms versus Lemmas in Long Texts. *PLoS ONE* **2015**, *10*, e0129031. [CrossRef]
51. Ferrer-i-Cancho, R.; Elvevåg, B. Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution. *PLoS ONE* **2010**, *5*, e9411. [CrossRef] [CrossRef] [PubMed]
52. Williams, J.; Lessard, P.R.; Desu, S.; Clark, E.; Bagrow, J.P.; Danforth, C.; Dodds, P. Zipf's law holds for phrases, not words. *Sci. Rep.* **2015**, *5*, 12209. [CrossRef] [CrossRef] [PubMed]
53. Corral, Á.; Serra, I. The Brevity Law as a Scaling Law, and a Possible Origin of Zipf's Law for Word Frequencies. *Entropy* **2020**, *22*, 224. [CrossRef] [PubMed] [CrossRef]
54. Ferrer-i-Cancho, R.; Solé, R.V. The Small-World of Human Language. *Proc. R. Soc. Lond. B* **2001**, *268*, 2261–2265. [CrossRef] [CrossRef]
55. Chen, Y.S.; Leimkuhler, F. A relationship between Lotka's Law, Bradford's Law, and Zipf's Law. *J. Am. Soc. Inf. Sci.* **1986**, *37*, 307–314. [CrossRef] [CrossRef]

56. Chen, Y.S.; Leimkuhler, F. Booth's law of word frequency. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 387–388. [CrossRef] [CrossRef]
57. Back, A.D.; Angus, D.; Wiles, J. Transitive Entropy—A Rank Ordered Approach for Natural Sequences. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 312–321. [CrossRef] [CrossRef]
58. Booth, A.D. A Law of occurrences for words of low frequency. *Inf. Control* **1967**, *10*, 386–393. [CrossRef] [CrossRef]
59. Montemurro, M.A. Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Phys. A Stat. Mech. Appl.* **2001**, *300*, 567–578. [CrossRef] [CrossRef]
60. Mandelbrot, B. *The Fractal Geometry of Nature*; W. H. Freeman: New York, NY, USA, 1983.
61. Taft, M.; Krebs-Lazendic, L. The role of orthographic syllable structure in assigning letters to their position in visual word recognition. *J. Mem. Lang.* **2013**, *68*, 85–97. [CrossRef] [CrossRef]
62. Fallows, D. Experimental evidence for English syllabification and syllable structure. *J. Linguist.* **1981**, *17*, 309–317. [CrossRef] [CrossRef]
63. Chetail, F.; Drabs, V.; Content, A. The role of consonant/vowel organization in perceptual discrimination. *J. Exp. Psychol. Learn. Mem. Cogn.* **2014**, *40 4*, 938–961. [CrossRef]
64. Port, R.; Dalby, J. Consonant/vowel ratio as a cue for voicing in English. *Atten. Percept. Psychophys.* **1982**, *32*, 141–152. [CrossRef] [PubMed] [CrossRef] [PubMed]
65. Davis, C.; Bowers, J. Contrasting five different theories of letter position coding: evidence from orthographic similarity effects. *J. Exp. Psychol. Hum. Percept. Perform.* **2006**, *32 3*, 535–557. [CrossRef]
66. Perry, C.; Ziegler, J.C.; Zorzi, M. A Computational and Empirical Investigation of Graphemes in Reading. *Cogn. Sci.* **2013**, *37*, 800–828. [CrossRef] [CrossRef]
67. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience: John Wiley & Sons: Hoboken, New Jersey, 2012.
68. Ferrer-i-Cancho, R.; Bentz, C.; Seguin, C. Optimal Coding and the Origins of Zipfian Laws. *J. Quant. Linguist.* **2020**, *0*, 1–30. [CrossRef] [CrossRef]
69. Chen, S.F.; Goodman, J. An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.* **1999**, *13*, 359–394. [CrossRef] [CrossRef]
70. Norvig, P. Natural Language Corpus Data. In *Beautiful Data*; Segaran, T., Hammerbacher, J., Eds.; O'Reilly: Sebastopol, CA, USA, 2009; pp. 219–242.
71. Norvig, P. English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU. 2020. Available online: https://norvig.com/mayzner.html (accessed on 17 November 2020).
72. Tanaka-Ishii, K.; Aihara, S. Computational Constancy Measures of Texts—Yule's K and Rényi's Entropy. *Comput. Linguist.* **2015**, *41*, 481–502. [CrossRef] [CrossRef]
73. Cover, T.; King, R.C. A convergent gambling estimate of the entropy of English. *IEEE Trans. Inf. Theory* **1978**, *24*, 413–421. [CrossRef] [CrossRef]
74. Brown, P.F.; Pietra, V.J.D.; Mercer, R.L.; Pietra, S.A.D.; Lai, J.C. An Estimate of an Upper Bound for the Entropy of English. *Comput. Linguist.* **1992**, *18*, 31–40.
75. Bentz, C.; Alikaniotis, D.; Cysouw, M.; Ferrer-i-Cancho, R. The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. *Entropy* **2017**, *19*, 275. [CrossRef] [CrossRef]
76. Debowski, L. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2021.
77. Kontoyiannis, I.; Algoet, P.; Suhov, Y.; Wyner, A. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inf. Theory* **1998**, *44*, 1319–1327. [CrossRef] [CrossRef]
78. Gao, Y.; Kontoyiannis, I.; Bienenstock, E. Estimating the Entropy of Binary Time Series: Methodology, Some Theory and a Simulation Study. *Entropy* **2008**, *10*, 71–99. [CrossRef] [CrossRef]
79. Takahira, R.; Tanaka-Ishii, K.; Debowski, L. Entropy Rate Estimates for Natural Language—A New Extrapolation of Compressed Large-Scale Corpora. *Entropy* **2016**, *18*, 364. [CrossRef] [CrossRef]
80. Kucera, H.; Francis, W.N. *Computational Analysis of Present-Day American English*; Brown University Press: Providence, RI, USA, 1967.