Zipf's law is not a consequence of the central limit theorem

G. Troll and P. beim Graben
Nichtlineare Dynamik, Universität Potsdam D-14415 Potsdam, Germany
(Received 23 April 1997)

It has been observed that the rank statistics of string frequencies of many symbolic systems (e.g., word frequencies of natural languages) follows Zipf's law in good approximation. We show that, contrary to claims in the literature, Zipf's law cannot be realized by the central limit theorem(s). The observation that a log-normal distribution of string frequencies yields an approximately Zipf-like rank statistics is actually misleading. Indeed, Zipf's law for the rank statistics is strictly equivalent to a power law distribution of frequencies. There are two natural ways to perform the infinite size limit for the vocabulary. The first one is the method of choice in the literature; it makes the upper word length bound tend to infinity and leads in the case of a multistate Bernoulli process via a central limit theorem to a log-normal frequency distribution. An alternative and for text samples actually better realizable way is to make the lower frequency bound tend to zero. This limit procedure leads to a power law distribution and hence to Zipf's law—at least for Bernoulli processes and to a very good approximation for natural languages where it passes the χ^2 test. For the Bernoulli case we will give a heuristic proof. [S1063-651X(98)07102-5]

PACS number(s): 05.40.+j, 87.10.+e

I. INTRODUCTION

This paper examines the meaning and the origin of Zipf's law [1] in the context of stochastic processes. First, we are going to state Zipf's law for symbolic systems.

Suppose we are given a finite or infinite string of symbols, such as a text in a natural language or a DNA sequence of a gene. Identify a set of constituent segments or building blocks. These may either suggest themselves in the specific context, such as proper linguistic words in the case of natural text or they may be just the finite strings in the general case. We will call them words anyway. Together they form the vocabulary. Next, we identify families of finite subvocabularies by introducing a parameter such as a fixed word length, an upper length bound or—as we will argue for in this paper—a lower frequency bound. Determine the multiset of word frequencies $\{p_i\}_{i\in I}$, i.e., we keep multiple instances of frequencies, and order its elements according to their decreasing size. Multiple instances of one frequency get consecutive ranks. The new index is called their rank. Zipf's law in the form given by Mandelbrot [2] now states that large enough samples close to the parameter limit obey approximately

$$p_r = \frac{B}{(A+r)^{\rho}}, \tag{1}$$

where r is the rank of the frequency p_r, A, B, and ρ are constants $(B,\rho>0)$. A suitable normalization condition leaves two free parameters, say A and ρ.

The main questions connected with Zipf's law concern its universality, its origin, and in particular its consequences such as short and long range correlations. It was claimed that Zipf's law can be found in many symbolic systems in linguistics, genetics, and even beyond symbolic systems in situations, where the rank ordering of quantities other than probabilities is examined (e.g., examples in economics). Examples treated in this paper are texts in natural languages, DNA sequences, and Bernoulli and Markov processes.

There have been many attempts to construct simple stochastic processes (Bernoulli and Markov) and also other models generating this law (e.g., [6,3,5,4,7]). See the Appendix for a short discussion. The arguments for a specific model were often largely phenomenological: by a least mean square fit the models just yield a reasonable approximation of Zipf's law over some intermediate range of ranks.

The aim of this paper is first to clarify the meaning of Zipf's law by transforming it to a distribution law. Doing this, one can replace least mean square fits by choosing the empirical mean and the empirical standard deviation as approximations of the mean and the standard deviation of the distribution function and interpret the free parameters in Eq. (1) as functions of the mean and the standard deviation. Furthermore, one can now distinguish more easily between distribution functions that yield a similar rank statistics. The second aim is to show how Zipf's law can be generated precisely and not only approximately by a broad class of stochastic processes.

It is clear from the algorithmic definition of the rank ordering given above that the word frequency distribution is equivalent to the rank statistics. What is done in rank statistics is just shift all word frequencies in such a way that they become equidistant, or in other words such that their distribution becomes uniform. This means analytically that we are given the Frobenius-Perron operator operating on the densities (or more generally the measures themselves) and we are looking for the point map $p_r \mapsto r$ associated with it. Details of the derivation can be found in Sec. II.

Using this transformation formula one finds that Zipf's law (1) is actually equivalent to an exponential distribution of the logarithmed word frequencies, which amounts to a power law distribution for the word frequencies themselves. This is sometimes called Zipf's second law. It is at odds with a normal distribution suggested in [7] and hence incompatible with the central limit theorem. We are going to show that for untruncated natural language texts the exponential distribution is a better approximation than the normal distribution over the whole range of data and is even a good

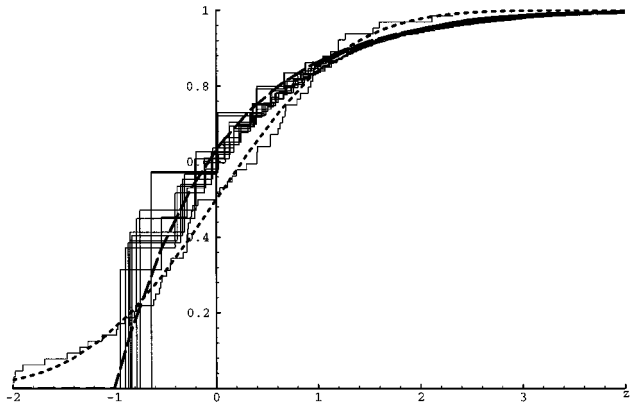

FIG. 1. The empirical distribution functions of the standardized logarithm of word frequencies z of several texts from German and English books are in the neighborhood of the standardized exponential distribution function (upper dashed curve), whereas the distribution function of codons of a yeast gene is much closer to the standardized normal distribution function (lower dotted curve).

approximation for small word frequencies. Figures 1 and 2 show the standardized word frequency distribution and the standardized rank statistics for several English and German texts and additionally for the codon (triplet) distribution for the DNA on a yeast chromosome.

In certain cases such as for the DNA the normal distribution is clearly the better approximation. Why? Using Bernoulli processes as the simplest models we will show in Sec. III that they can actually realize two different limit theorems. In the case covered by the central limit theorem the infinite size limit for the ensemble of word frequencies is performed by making the word length bound L tend to infinity. However, there is an alternative way. Instead of parametrizing by the word length one can introduce a lower frequency bound ϵ as a cutoff. We show numerically and prove heuristically that by taking $\ln \epsilon \to -\infty$ in the Bernoulli process one does not obtain the normal distribution but the exponential one as a limit. Performing the limit in this way seems to us to be more suitable for finite samples of Bernoulli language models or actual texts of natural languages where smaller word frequencies are harder to find although in the case of the $L \to \infty$ limit they should be included already for moderate L because the letter frequencies vary considerably.

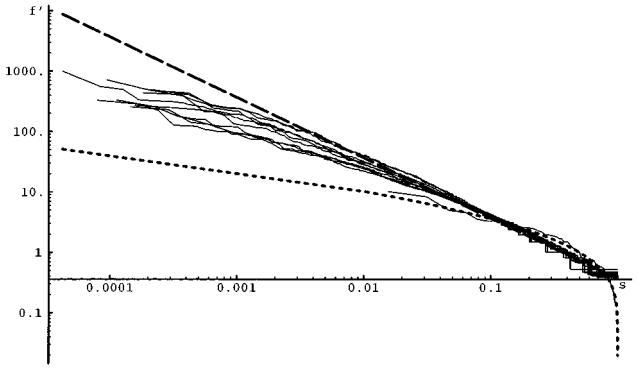

FIG. 2. The empirical rank statistics $f'(s) = \exp z$ of the data of Fig. 1 as a function of the normalized rank s. The exponential distribution appears here as the dashed straight line (Zipf's law), the normal distribution as the lower dotted curve.

II. ZIPF'S LAW FOR DISTRIBUTIONS

A. From word frequency to rank statistics

In this section we are going to study the relationship between the distribution of word frequencies to the rank ordered frequencies by means of the Frobenius-Perron operator. Our general setting is that of a stationary information source given by a stochastic process (A^N, F_A, P), where A^N is the set of sequences over a finite set A, called the alphabet and F_A is the Borel field of subsets of A^N, which is determined by the cylinder sets ${}_n[a_n \cdots a_m]_m = \{s \in A^N; s_i = a_i \text{ for } n \leq i \leq m\}$; we assume that P is a shift-invariant (i.e., stationary) probability measure on F_A, so that we can identify the cylinder sets with finite strings: ${}_n[a_n \cdots a_m]_m = a_n \cdots a_m$. Let W now be a subset of strings, called the total vocabulary.

The simplest case is that W is just the set of all finite strings, i.e., the string vocabulary, but we are more interested in the case where it is a true subset. For natural languages the natural choice of this subset is formed by linguistic words and the information source is given by a finite string T, called the text. We are only interested in the values of P on W, which we define by the relative frequencies of the words in the text T. For practical purposes this means that the text must be long enough to permit a reliable estimate of P and short enough to yield approximately stationary word frequencies.

Let (W_i) be a nested sequence of finite subvocabularies (the ith step vocabularies) whose union is W. If W is the string vocabulary we choose the string length as a parameter. In the other case we define two special types of finite subvocabularies parametrized by the upper word length bound L and by the lower word probability bound ϵ, respectively: let $W_{l \leq L}$ be the set of words in W of length $l \leq L$ and $W_{p \geq \epsilon}$ the set of words with probabilities $p \geq \epsilon$. We study the families $(W_i) = (W_{l \leq L})_L$ and $(W_i) = (W_{p \geq \epsilon})_\epsilon$.

Denote by U_i the multiset (i.e., keeping all instances of the same element) of relative frequencies of words in W_i, i.e., U_i has the same cardinality as W_i, which is the vocabulary size denoted by $\#W_i$. By ''#'' we denote the number of elements in the set following this symbol.

Example II.1: The first information source is determined by the Luther bible, whose length is about 5×10^6 characters; A is the set of small letters of the German alphabet together with the blank and punctuation symbols; we define two different total vocabularies: $W^{(1)} = W^{\text{ling}}$ is the linguistic vocabulary consisting of all the bible's proper names and German words ($\#W^{\text{ling}} = 23679$) and $W^{(2)}$ is the string vocabulary, i.e., the set of all substrings of the Luther bible; $P^{(1)}$ and $P^{(2)}$ are determined by the relative frequencies of words in $W^{(1)}$, $W^{(2)}$, respectively.

Example II.2: Take the DNA string of chromosome III of the yeast Saccharomyces cerevisiae strain S288C with length ca. 3.2×10^5 over the alphabet $A = \{A, G, C, T\}$ of DNA

bases; W is the string vocabulary; there is a special vocabulary $W_{l=3}=A^3$ consisting of what is called codons (base triplets) in genetics; $\#W_{lci}=4^i$, i.e., 64 for the codons.

Example II.3 (Perline 96): Here we take the $(K+1)$-state Bernoulli process; the alphabet is the state set $A=\{L_1,\ldots,L_K,\square\}$, $K\geqslant 2$, ''$\square$'' is the space character, with probabilities $a=(a_1,\ldots,a_{K+1})$, $\Sigma a_i=1$, $\max a_i=a_{K+1}$, which is supposed to be nongenerate in the sense that $\{a_1,\ldots,a_K\}$ contains at least 2 (different) elements; again $W^{(1)}$ is the set of those substrings that are delimited by the space character and is supposed to simulate the linguistic vocabulary of a text in a natural language; $W^{(2)}$ is the substring vocabulary, where we might ignore the space character.

Example II.4 (Kanter, Kessler 95): The stochastic process taken here is a Markov process, its state set is $A=\{0,1,\ldots,N-1\}$, $N=2^L$; each state $m\in A$ is connected to just two states $m_0=2m \bmod N$, $m_1=2m \bmod N+1$. The transition matrix S_1 is determined by giving these 2 transitions probabilities x, $1-x$, respectively, independent of m, whereas transition matrix S_2 does the same with probabilities $1-x,x$, respectively: The actual trial chooses S_1 and S_2 with probabilities (bias) B and $1-B$, respectively.

Next we define a probability space on U_i. We introduce the random variable $Y_i:W_i\to U_i$, which associates to each word w its probability p; take as an σ algebra on U_i the power set and as a measure the counting measure. In the following we are more interested in the random variable $\ln Y_i$, which we standardize introducing $Z_i=(\ln Y_i-m_i)/\sigma_i:W_i\to V_i$, where m_i is the mean and σ_i the standard deviation of $\ln Y_i$, $V_i=Z_i(W_i)$ as multiset.

Let M_i be the distribution function and μ_i the probability distribution of Z_i, i.e., we have

$$M_i(x)=\mu_i(\{Z_i\leqslant x\})=\frac{\#\{q\in V_i,q\leqslant x\}}{\#V_i},\quad x\in\mathbb{R}. \tag{2}$$

We assume now that the distribution functions M_i converge pointwise to a distribution function M with probability distribution μ:

$$\forall x\in\mathbb{R}: M_i(x)\xrightarrow{i\to\infty} M(x). \tag{3}$$

The limit distribution μ can only then be continuous if the vocabulary size $\#W_i$ tends to infinity.

Next we are going to introduce rank statistics. The random variable Z_i is realized by a sequence $(q_j)_{j\in Ji}$ of frequencies. Usually, the rank map is regarded as a permutation on the index set J_i which orders the sequence $(q_j)_{j\in J_i}$ according to size. For our purposes it is more convenient to introduce the normalized rank $s=r/\#J_i$ as the values of a monotonically decreasing *rank map* S_i into the interval $[0,1]$, which satisfies $q_j\leqslant q_k$ iff $s_j=S_i(q_j)\geqslant S_i(q_k)=s_k$, so that q_j is the r_jth largest value of $(q_j)_{J_i}$. We do this because we want to extend the rank map to the real line where it is more amenable to analytic calculations. The problem with this definition of the rank map is that we have defined it on a sequence or equivalently on a multiset $W_i=\{q_j\}_{j\in J_i}$. But there is an easy solution. We remove the degeneracy of multiple values in the multiset W_i by perturbing them by an amount that is small relative to the minimal nonvanishing spacing of neighboring values in W_i. The precise form of the perturbation is arbitrary. Essential is only that we arrive at a genuine set $\widetilde{W}_i$ and corresponding $\widetilde{Z}_i$ with distribution function $\widetilde{M}_i$. Alternatively, if we do not have additional symmetries enforcing degeneracy, we may interpret multiple occurrences of a frequency in V_i as a finite size effect, i.e., we interpret the multiset V_i as an imprecise measurement of the limit frequency set V where all frequencies differ. The inverse rank map S_i^{-1} exists in any case.

Now, suppose $S_i:\widetilde{V}_i\to[0,1]$ is a rank map of $\widetilde{Z}_i$. The rank ordered log-frequency curve is given by $\{(S_i(x),x);x\in\widetilde{V}_i\}$. Let ν_i be the counting measure on $\Delta_i:=\{1,\ldots,\#W_i\}/\#W_i$, i.e., $\nu_i(\{s\})=1/\#W_i$ for $s\in\Delta_i$ and 0 on all other singletons of $[0,1]$. Its distribution function is $N_i(s)=\Delta_i(s)$, $s\in[0,1]$, where $\Delta_i(s)$ is the closest element $\leqslant s$ in $\Delta_i\cup\{0\}$. Once $\widetilde{V}_i$ is fixed, the rank map is determined by its action on measures: its Frobenius-Perron operator $\mathcal{P}_{S_i}$ transforms the measure $\widetilde{\mu}_i$ to the counting measure ν_i:

$$\mathcal{P}_{S_i}\widetilde{\mu}_i(A)=\nu_i(A)\quad\text{for each set } A\subset\Delta_i, \tag{4}$$

where $\mathcal{P}_{S_i}\widetilde{\mu}_i(A)=\widetilde{\mu}_i(S_i^{-1}(A))$ by definition of the Frobenius-Perron operator for measures. We extend S_i^{-1} monotonically decreasing but otherwise arbitrarily to $[0,1]$. Hence, we get for $s\in\Delta_i$ and $A=[s,1]$ the equation $\widetilde{M}_i(S_i^{-1}(s))=1-N_i(s)$ or $\widetilde{M}_i=1-S_i$ on Δ_i.

Now we let $i\to\infty$. Then the counting distribution N_i tends to the continuous uniform distribution on $[0,1]$, whose distribution function is the identity map. Therefore

$$S_i\to S=1-M \tag{5}$$

pointwise because $\widetilde{M}_i\to M$. If the limit distribution M is continuous it is surjective, hence the right inverse exists and we also have the existence of the inverse rank map

$$S^{-1}(s)=M^{-1}(1-s)\quad\text{for } s\in\,]0,1[. \tag{6}$$

Remark II.1: Observe that this formula remains valid also for not standardized distributions M_n and M. Furthermore, Eq. (6) is equivalent to

$$\mathcal{P}_S\mu(A)=\nu(A) \tag{7}$$

for each Borel set $A\subset[0,1]$. If the limit distribution M is continuous one knows that the convergence of M_n is automatically uniform. In this case we also have $S_n^{-1}\to S^{-1}$ pointwise. In the following figures, both the rank maps $f'(s)=\exp z(s)$ and the distribution function $M_i(z)$ are plotted. One sees readily that the uniform convergence of distributions contrasts with only pointwise convergence of rank maps.

Definition II.1: The generalized rank map for a distribution μ is given by Eq. (7) or equivalently (6).

Thus, we have shown the following:

Lemma II.1: For the sequence of random variables Z_i which converges in distribution the rank maps converge pointwise to the rank map of the limit distribution.

B. What distribution generates Zipf's law?

We show here that Zipf's law for the rank statistics is equivalent to a power law of the word frequency distribution. Let X be a real valued random variable (playing the role of $\ln Y_i$ above) with mean m, standard deviation σ, and distribution function E_X. Furthermore, suppose the inverse rank map $f:=S^{-1}_{\exp X}$ of the random variable e^X obeys Zipf's law with parameters $A \geqslant -1$, $B>0$, $\rho>0$, i.e.,

$$S^{-1}_{\exp X}(s)=\frac{B}{(s+A)^{\rho}}. \tag{8}$$

What is the form of E_X? The logarithm of the frequencies obey Eq. (6):

$$x:=\ln S^{-1}_{\exp}(s)=\ln B-\rho \ln(s+A)=E_X^{-1}(1-s). \tag{9}$$

Solving the first part of this double equation for s and setting the result into the second part yields $E_X(x)=1+A-B^{1/\rho}\exp(-x/\rho)$ for $x>C$, 0 otherwise, where $C\in\mathbb{R}\cup\{-\infty\}$ is determined as the left border of the support of E_X, i.e., $E_X(x)=0$ for all $x\leqslant C$. The requirement $E_X(\infty)=1$ implies $A=0$ and $E_X(C)=0$ amounts to $B=e^C$; calculating the mean and the standard deviation yields $C=m-\sigma$ and $\rho=\sigma$, and consequently

$$E_X(x)=\begin{cases}1-e^{-(x-m+\sigma)/\sigma} & \text{for } x>m-\sigma\\ 0 \quad \text{otherwise}\end{cases} \tag{10}$$

for the distribution of X. For the random variable e^X with realization $\xi=e^x$ we receive the power function $E_{\exp X}(\xi)=1-e^{(m-\sigma)/\sigma}\xi^{-1/\sigma}$ for $\xi\geqslant e^{m-\sigma}$ and 0 otherwise. Its density is $e_{\exp X}(\xi)=E'_{\exp X}(\xi)=(1/\sigma)e^{(m-\sigma)/\sigma}\xi^{-1-1/\sigma}$ for $\xi\geqslant e^{m-\sigma}$, and 0 otherwise. Using the moments of X Zipf's law can be written as

$$S^{-1}_{\exp X}(s)=\frac{\exp(m-\sigma)}{s^{\sigma}}, \quad s\in\,]0,1]. \tag{11}$$

Now assume conversely that $\ln X$ is exponentially distributed [Eq. (10)]. Then $\ln S^{-1}_{\exp X}(s)=S_X^{-1}(s)=E_X^{-1}(1-s)=-\sigma\ln s+m-\sigma$, i.e., Zipf's law. In a log-log plot Zipf's rank curve will be a line: with $t=\ln s$, $\zeta(t):=\ln f(e^t)=-\sigma t+m-\sigma$ where $t\in\,]-\infty,0]$. The limit case $\rho=0$ is equivalent to a uniform distribution.

Now, take the log-frequency random variable $\ln Y_i$ for X. Zipf's hypothesis is that $S^{-1}_{Y_i}$ obeys approximately Zipf's law. As $e^{Z_i}=Y_i^{1/\sigma_i}\exp(-m_i/\sigma_i)$ we find that the limit distribution function of Z_i is exponential, i.e., $M(x)=1-e^{-(x+1)}$ for $x>-1,0$ otherwise if and only if its limit rank map obeys Zipf's law in its standardized form

$$\lim S^{-1}_{\exp Z_i}(s)=S^{-1}_{\exp Z}(s)=\frac{1}{es}, \quad s\in\,]0,1]. \tag{12}$$

This is of course also true for discrete random variables.

Theorem II.2: The empirical rank statistics of word frequencies converges to Zipf's law if the empirical distribution of log-frequencies converges in distribution to the exponential distribution. The convergence of distribution functions is automatically uniform.

C. What rank curve is generated by a process obeying the central limit theorem?

Suppose the empirical distribution function converges pointwise to the normal distribution:

$$\forall x\in\mathbb{R}: M_i(x)\xrightarrow{i\to\infty} N_{0,1}(x)=\int_{-\infty}^{x}e^{-y^2/2}dy. \tag{13}$$

Examples II.3 and II.4 satisfy this condition. For example II.3 this does not follow directly from the central limit theorem because the word length is not fixed. Taking the word length as another random variable Perline shows in [7] by using a variant of the central limit theorem (Anscombe's theorem, cf. theorem 3.1 in [8]) that the word frequencies generated by the nondegenerate Bernoulli process of example II.3 converge in distribution to the normal distribution:

$$M_{l\leqslant L}\to N_{0,1} \quad \text{for} \quad L\to\infty. \tag{14}$$

The stationary probability densities of the Markov process of example II.4 also seem to be approximately normally distributed if the number of states is large enough. For an important special case $B=0$ of this Markov process this is straightforward: The stationary probabilities are given in [4] as $p_j=q^j/(1+q)^L$ for all words containing j 1's and $N=2^L$. However, from this one sees immediately that the L logarithm of the probabilities are equidistant and occur $\binom{L}{j}$ times, so that the probabilities themselves are binomially distributed. For L tending to infinity the binomial distribution converges to the normal distribution by the theorem of de Moivre and Laplace.

We know already that the log-normal distribution does not imply Zipf's law analytically. This leads to two questions: (i) Does the log-normal distribution yield at least a good approximation of Zipf's law? (ii) Does the log-normal distribution yield a good approximation of the empirical distribution function of word frequencies?

1. Does the log-normal distribution yield a good approximation to Zipf's law?

The rank statistics for the normal distribution is given by Eq. (6) as

$$\ln f(s)=N_{0,1}^{-1}(1-s) \quad \text{for } s\in\,]0,1[. \tag{15}$$

By Eq. (12) a good approximation to Zipf's standardized law requires that $\tau:=\zeta(t)=N_{0,1}^{-1}(1-e^t)\approx -t-1$, which is satisfied if $\zeta'=-1$ and for $\zeta(t\approx 0)\approx -1$. Observe that $t\approx 0$ corresponds to $s\approx 1$, that is to a high rank. However, as $\zeta^{-1}(\tau)=\ln[1-N_{0,1}(\tau)]$, we get

$$\zeta'(t)=1/(\zeta^{-1})'(\tau)=-\frac{1-N_{0,1}(\tau)}{N'_{0,1}(\tau)}. \tag{16}$$

Using the error function one can find an asymptotic expansion of this expression for large τ, which corresponds to large word frequencies: $\zeta'(\tau)\approx-(1/\tau)[1-(1/\tau^2)\pm\cdots]$. However, this is only in the neighborhood of -1 for $\tau^3-\tau^2+1=0$, i.e., $\tau\approx -0.75488$, which is not in the asymptotic region. Of course, it is possible to approximate

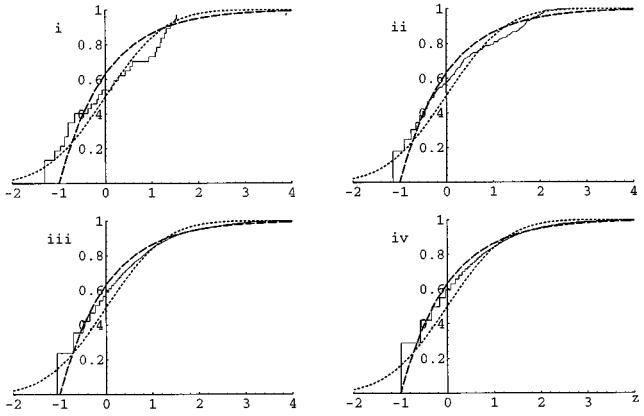

FIG. 3. The empirical distribution functions of the standardized logarithm of word frequencies z of the Luther bible. Admitted are only linguistic words of length (i) $L \leqslant 2$, (ii) $L \leqslant 3$, (iii) $L \leqslant 5$, (iv) $L \leqslant 7$. The upper dashed curve is the standardized exponential, the lower dotted curve the normal distribution function.

Zipf's law in a certain region by a least mean square fit. This has been the standard procedure in the literature but the selection of the region is arbitrary and the fit will lead to even larger deviations in other regions.

2. Does the log-normal distribution yield a good approximation of the empirical distribution function of word frequencies?

It is usually argued that the rank statistics of a broad range of samples obeys Zipf's law, i.e., the rank statistics is approximately linear in a log-log plot. So it is claimed, for instance, by Kanter and Kessler that both the bible and the codons of the yeast DNA satisfy Zipf's law. However, in the empirical distribution functions the qualitative differences are immediately apparent. Figures 1 and 2 show the word frequency distribution and rank statistics for several English and German texts (by Martin Luther, Immanuel Kant, Charles Dickens, Frank Baum, and Charles Darwin) and additionally the codon (triplet) distribution for the DNA on chromosome III of the yeast Saccharomyces cerevisiae strain S288C which was sequenced at Manchester Biotechnology Centre, UMIST, UK, 1992. All texts are clearly rather exponentially than normally distributed (in particular for values corresponding to larger word frequencies) whereas the DNA is well approximated by a normal distribution over the whole range.

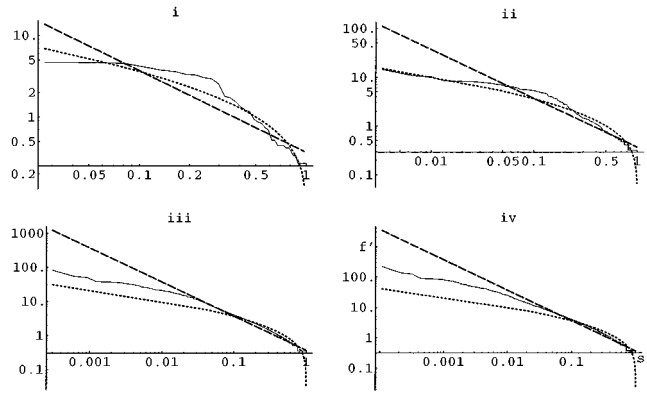

FIG. 4. The same as Fig. 3 for the rank statistics $f' = \exp z$.

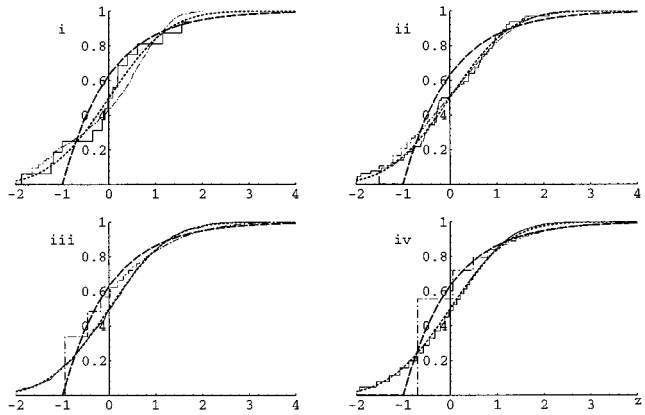

FIG. 5. The empirical distribution functions of the standardized logarithm of word frequencies z of the Luther bible (dash-dotted line) and of the yeast DNA (continuous line). Admitted are only strings (blanks are canceled) of fixed length (i) $L = 2$, (ii) $L = 3$, (iii) $L = 5$, (iv) $L = 7$.

Still, the answer is not yet complete, because we have sampled all words, whereas the central limit of Eq. (14) requires neglecting words of lengths above a cutoff parameter L. It is reasonable to suspect that the statistics for longer word lengths L eventually deteriorates in a given finite text. Therefore one might get a better approximation to the normal distribution for intermediate L. Figures 3, 4, 5, and 6 show snapshots for word lengths $L = 2,3,5,7$ for the linguistic vocabulary of the Luther bible and the string vocabulary of the Luther bible together with the yeast DNA, respectively. Whereas for the linguistic vocabulary the normal distribution can be seen even for $L = 2,3$ only with some imagination, it is much more visible for $L = 2,3$ in the string vocabulary of the Luther bible as a transient state. For larger $L = 5,6,7$ the exponential distribution takes over again. This is in contrast to the central limit for DNA substrings, where one sees very distinctly the normal distribution as high L limit.

There is no point going much beyond $L = 7$ because the word number $\#W(l)$ as a function of word length l reaches its maximum at $l = 7$ for both the linguistic words and the substrings of the Luther bible, so that for $l > 7$ the topological entropy sequence $l \mapsto \ln \#W(l)/l$ deviates qualitatively from a Bernoulli process.

Of course, there is no *a priori* reason why a Bernoulli process (example II.3) should be a good model for natural

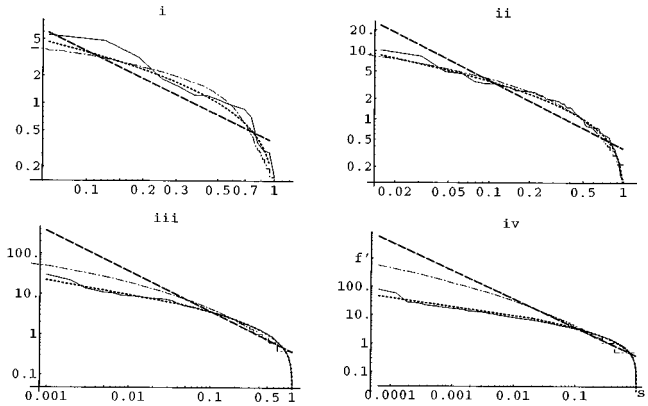

FIG. 6. The same as Fig. 5 for the rank statistics $f' = \exp z$.

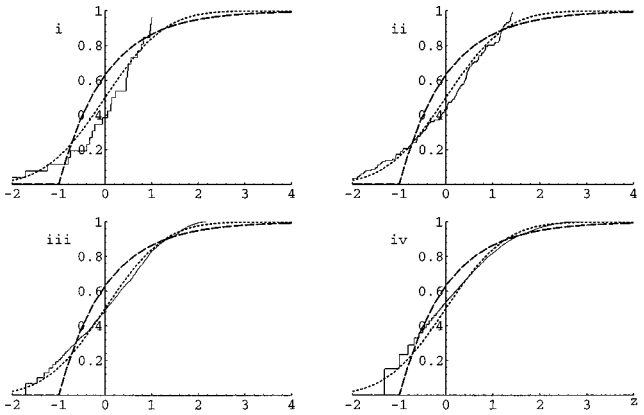

FIG. 7. The empirical distribution functions of the standardized logarithm of word frequencies z of the Luther bible recoded in a phonetic 5 letter alphabet. Admitted are only strings (blanks are canceled) of fixed length (i) $L=2$, (ii) $L=3$, (iii) $L=5$, (iv) $L=7$.

languages whose letter probabilities are obviously not independent. Choosing such a model and taking the word length limit $L\to\infty$ results only therefore in the normal distribution because those short words that are made up completely of rare letters produce both the smallest word frequencies and small numbers of letter permutations and thus generate the left branch of the normal distribution. However, it is precisely these words that are missing in natural texts. For instance, the rarest 4 letters in English are ''z,'' ''q,'' ''x,'' and ''j.'' They are a few hundred times rarer than the most common letter ''e.'' The Random House Unabridged, one of the largest unabridged dictionaries of American English, contains only 21 words with 2 ''x,'' 1 with 3 ''x,'' 1 (a hyphenated proper name) which contains all three letters ''z,'' ''q,'' ''j.'' These vocabulary gaps are probably the reason why the central limit also (see next section) yields exponential rather than normal distributions for natural texts of the examined sizes of up to a few Mbytes. The difference of DNA sequences to natural languages on the level of the Bernoulli processes is that their letter (the bases in genetics) frequencies vary much less and word gaps do not occur.

One way to test this hypothesis is to recode the text by choosing a coarser code with a more balanced letter frequency than the Latin alphabet. By this method the gaps of rare letter words can be closed without changing the distribution of word lengths. In Figs. 7 and 8 we took a 5 letter phonetic encoding. As predicted, the convergence to a normal distribution is now much better.

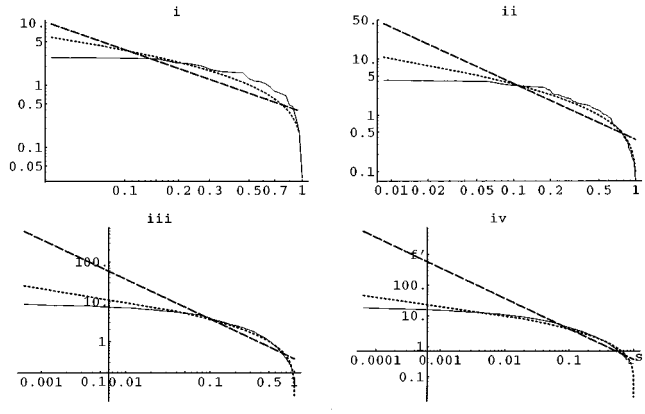

FIG. 8. The same as Fig. 7 for the rank statistics $f'=\exp z$.

III. AN ALTERNATIVE WAY TO PERFORM THE INFINITE SIZE LIMIT

In this section we ask the question of whether Zipf's law can already be produced by a Bernoulli process. This question is important because if the answer is yes, then one cannot deduce the existence of any correlations for a stochastic process which obeys Zipf's law. At first glance this question seems to have been answered to the negative, because the Bernoulli process leads to the log-normal distribution, which as we have argued can be numerically and analytically distinguished from Zipf' law. However, as we have already stressed in the introduction there is an alternative way to perform the limit in 3.

Instead of parametrizing by the upper bound of word lengths we choose a lower bound ϵ of word frequencies as cutoff parameter. That is we neglect words that are rarer than a fixed frequency parameter and study the subvocabularies $(W_i)=(W_{p\geq\epsilon})_\epsilon$ for $\epsilon\to 0$ or rather for technical reasons $(W_i)=(W_{\ln p\geq f})$ for the cutoff parameter $f\to-\infty$. What are the consequences for a Bernoulli process?

Take example II.3 with state set $A=\{L_1,L_2,\square\}$ and probabilities $a=(a_1,a_2,a_3)$, $\Sigma a_i=1$, $\max(a_i)=a_3$. It is easy to prove that the distribution of log frequencies of words in $(W_{\ln p\geq f})$ cannot converge to a normal distribution for $f\to-\infty$. First, assume that the log probabilities $\ln a_1$ and $\ln a_2$ are incommensurable (i.e., $\ln a_1/\ln a_2$ is irrational). Then for each normalized log probability f (normalized by subtracting $\ln a_3$) there is at most one pair (n_1,n_2) of natural numbers, such that $\varphi=n_1\ln a_1+n_2\ln a_2$ describing the normalized log probability of all words with n_1 symbols L_1 and n_2 symbols L_2 and ending by definition with the blank symbol $\square$. Their number $\#\varphi$, which we call degeneracy of φ, is given by the binomial coefficient.

$$\#\varphi=\#\varphi(n_1,n_2)=\binom{n_1+n_2}{n_1}=:\mathrm{Binom}(n_1+n_2,n_1). \tag{17}$$

If $\ln a_1$ and $\ln a_2$ are commensurable then

$$\#\varphi=\Sigma_{f=n_1\ln a_1+n_2\ln a_2}\mathrm{Binom}(n_1+n_2,n_1)\,.$$

The limit case $a_1=a_2$ leads to Mandelbrot's model (see Appendix) where all words of the same length get the same probability. This will result in an exponential distribution of φ for any f. If on the other hand $a_1<a_2$ then $f\mapsto\max\{\#\varphi;\varphi\in[f,f+\ln a_2]\}$ is monotonically decreasing. This means that on the scale of $\ln a_2$ the degeneracy $\#\varphi$ is a monotonically decreasing function of φ because one can add another symbol L_1 (there is no string length bound now) to a string that realizes the global maximum up to some f without moving beyond $f-\ln a_2$. As for small enough f it is possible to swap symbols L_2 by symbols L_1 without changing f by more than $\ln a_2$ the degeneracy $\#\varphi$ will diverge to infinity for $\varphi\to-\infty$. Therefore the standardized distribution function M_f of log frequencies cannot have an inflection point and thus does not converge to the normal distribution for cutoff

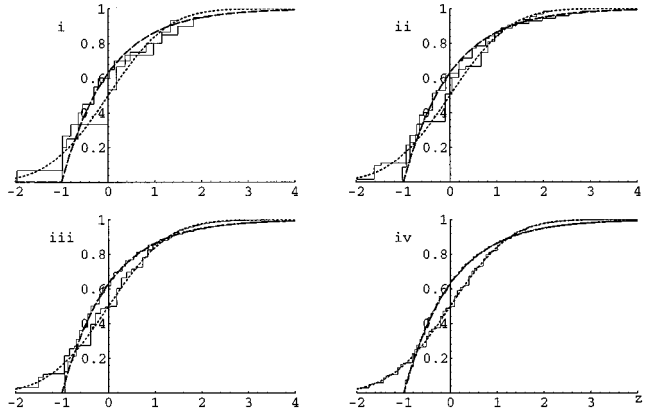

FIG. 9. The empirical distribution functions of the standardized logarithm of word frequencies z of a three state Bernoulli process. Admitted are those blank separated words obeying a word bound (lower curves) and those obeying a frequency bound (upper curves), respectively: (i) $L \leqslant 3$, $\ln p \geqslant -10$, (ii) $L \leqslant 5$, $\ln p \geqslant -25$, (iii) $L \leqslant 10$, $\ln p \geqslant -50$, (iv) $L \leqslant 40$, $\ln p \geqslant -100$.

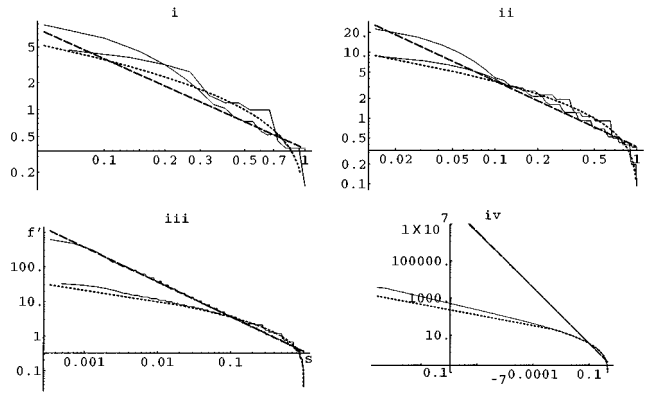

FIG. 10. The same as Fig. 9 for the rank statistics $f' = \exp z$.

parameter $f \to -\infty$. If $\ln a_1$ and $\ln a_2$ are incommensurable then M_f does not converge to a staircase function. One can see that as follows. For an arbitrarily small gap size $\delta > 0$ there is a $N_\delta > 0$ s.t. for any $n \geqslant N_\delta$ the frequencies $\varphi = n_1 \ln a_1 + n_2 \ln a_2$ will fill the interval $[n, n+1] \log a_2$ leaving only gaps of length $\leqslant \delta$. Only on these gaps is M_f constant. As $M_f(N_\delta \ln a_2)$ will be arbitrarily close to 1 if the cutoff f is small enough, the limit of M_f cannot contain a staircase.

Lemma III.1: If the $(K+1)$-state Bernoulli process of example II.3 contains at least two incommensurable log-probabilities of nonspace characters then the standardized distribution function M_f converges neither to the normal distribution nor to a staircase for $f \to -\infty$.

Conjecture III.2: If the $(K+1)$-state Bernoulli process of example II.3 contains at least two incommensurable log probabilities of nonspace characters then the standardized distribution function M_f converges to the exponential distribution function for $f \to -\infty$.

We have only the following idea of a heuristic proof for $K = 2$:

As above one sees that one can restrict oneself to an interval $[f, N \ln a_2]$, where $|N \ln a_2|$ can be taken arbitrarily large if $|f|$ is large enough, because the complement $[N \ln a_2, 0]$ has arbitrarily small μ_f measure. Points $\varphi \in [f, N \ln a_2]$ have a representation $\varphi = n_1 \ln a_1 + n_2 \ln a_2$ with large $n = n_1 + n_2$. By swapping steps $\ln a_1$ by steps $\ln a_2$ (possible because of incommensurability) one can find in a small neighborhood of φ a $\varphi' = n_1' \ln a_1 + n_2' \ln a_2$ with high weight $\mathrm{Binom}(n_1' + n_2', n_1')$. Thus, for any frequency φ its degeneracy $\#\varphi$ is locally determined by a dominant coefficient $\mathrm{Binom}(l,k)$. Moving in steps of size $-\ln a_2$ towards $\varphi = 0$ will result in either the dominant coefficient changing $\mathrm{Binom}(l,k) \to \mathrm{Binom}(l-1,k) = \mathrm{Binom}(l,k)(l-k)/l$, which is realized by just canceling one of the more common symbols L_2 and keeping the number of the rarer symbols L_1 constant or by swapping m of the k rarer symbols into $m' > m$ more common ones: $\mathrm{Binom}(l,k) \to \mathrm{Binom}(l+m', k-m)$. In the first case the factor λ of dominant coefficients per step $-\ln a_2$ is roughly constant but decreasing with every step. The second case, however, will increase λ again. The total effect is that already for moderate f and subsequent standardization λ is nearly constant for large l, whereas for smaller l the accumulation of the larger l terms dominates so that an approximate exponential distribution function results. Figures 9 and 10 show the numerical convergence of the central limit $L \to \infty$ to the normal distribution and of the frequency bound limit $f \to -\infty$ to the exponential distribution. The three-state Bernoulli process taken there was determined by $\ln a_1 / \ln a_2 = 2\pi$.

For a text of a natural language Figs. 11 and 12 show that the frequency bound limit $f \to -\infty$ also leads to a very good approximation of the exponential distribution, i.e., to Zipf's law. The closest mean square distance is reached for an f

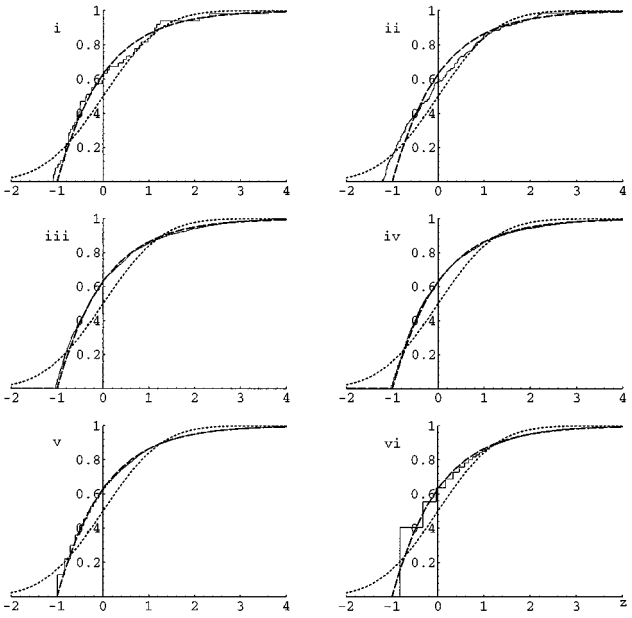

FIG. 11. The empirical distribution functions of the standardized logarithm of word frequencies z of the Luther bible. Admitted are only linguistic words with base frequencies p (normalized to the most frequent word) satisfying (i) $p \geqslant 0.05$, (ii) $p \geqslant 0.01$, (iii) $p \geqslant 0.001$, (iv) $p \geqslant 6.4 \times 10^{-4}$ (i.e., words occurring $\geqslant 30$ times), (v) $p \geqslant 8.8 \times 10^{-5}$ (i.e., occurring $\geqslant 5$ times), (vi) $p \geqslant 0$.

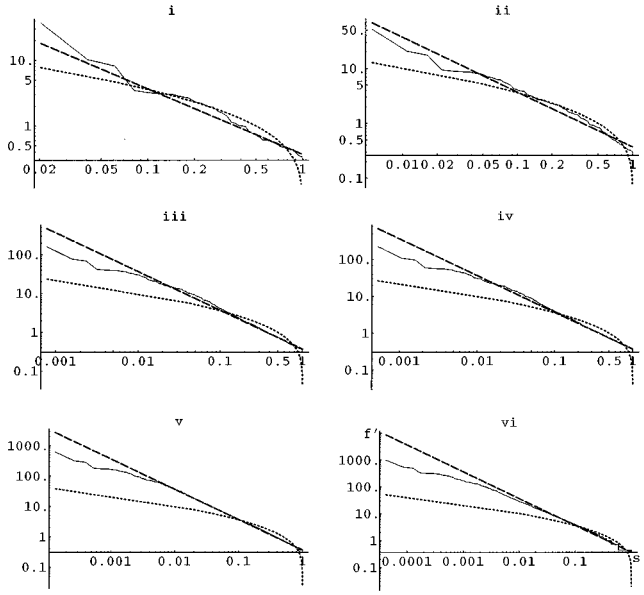

FIG. 12. The same as Fig. 11 for the rank statistics $f' = \mathrm{expz}$.

corresponding to at least $f_{\mathrm{abs}} = 30$ occurrences (part iv of the figures), the closest maximum distance for at least 5 occurrences (part v). For still lower frequencies the sequence moves away from Zipf's law presumably because the statistics deteriorates for words that are too rare relative to the given text. The χ^2 test for the exponential distribution yields at $f_{\mathrm{abs}} = 30$ the value $\chi^2 = 196$. As $P(\chi^2 \geqslant 196) \approx 10\%$ the hypothesis of an exponential distribution cannot be statistically rejected at a significance level of 5%, whereas the normal distribution can always be rejected with corresponding $P \approx 0$.

This shows that the frequency bound limit leading to Zipf's law is much better realized in natural texts than the central limit leading to a normal distribution. We think the reason for this is that the gaps of missing rare letter words that perturb the central limit are much less critical for the frequency bound limit because their number is overwhelmed by the number of long frequent letter words with similar frequency.

IV. CONCLUSIONS

We have shown that two clearly distinguishable infinite size limits play a role in the frequency statistics of symbolic systems: the central limit leading typically to a normal distribution and the frequency limit leading typically to an exponential distribution. Zipf's law is realized by the frequency limit and not by the central limit as was claimed in the literature. Nevertheless, Zipf's law is only a statistical phenomenon, which appears already in a Bernoulli process. Therefore it does not reflect any dynamically nontrivial properties of the underlying system. In particular, it does not require long range correlations or any correlations at all. Moreover, it does not require any pruning (forbidden letter sequences) or is necessarily destroyed by pruning.

ACKNOWLEDGMENTS

We would like to thank J. Kurths and C. Scheffzcyk for having put our attention to some of the literature used in this article. The authors acknowledge support by the Deutsche Forschungsgemeinschaft within the Innovationskolleg Formale Modelle kognitiver Komplexität, Potsdam. G. T. was additionally supported by the Max-Planck-Gesellschaft.

APPENDIX: SOME MODELS FOR ZIPF'S LAW STUDIED IN THE LITERATURE

The simplest model for Zipf's law yielding a power law distribution of word frequencies has been invented by Mandelbrot (cf. [3,5]). One of his numerous suggestions was a degenerated Bernoulli process with the same probability p for all letters and a different probability q for the space symbol. As all words of length i get the same probability $p^i q$ this leads to a geometrical distribution of word lengths. The counting density is just the word number $K^i = d(p^i q)$ where K denotes the cardinality of the alphabet. Eliminating i in $x = p^i q$ one obtains a power law $d(x) = (x/q)^{\ln K/\ln p}$ of word probabilities.

Other models of Mandelbrot generating the same power law distribution for word frequencies simulate the evolution of the vocabulary of a language in time such as Markov processes operating on word length and on the frequency of word use. Furthermore, Mandelbrot has given a model independent interpretation of Zipf's law in terms of coding theory: he asked how the word probabilities must be distributed in order to maximize the Shannon entropy of a message under the constraints of normalization and fixed averaged coding costs [6,3]. Assuming that coding costs of words are proportional to the word length, the thermodynamical formalism provides a canonical ensemble for word probabilities. However, this description leads to the first model of a degenerated Bernoulli process by interpreting the reciprocal of the partition function as probability q of the space symbol and the Shannon information of a genuine letter as the reciprocal ''temperature of discourse.''

In [7] Perline proves by using a variant of the central limit theorem (Anscombe's theorem) that a nontrivial m-state Bernoulli process, i.e., one that has more than one letter probability, generates log-normally distributed word frequencies if one selects only words below a fixed word length L and performs the limit by taking $L \to \infty$ (for words of the same length L this result is due to [3]). Instead of examining the rank statistics induced by the log-normal distribution directly, he addresses the problem of retrieving the rank statistics of a Bernoulli process as a ''broken stick'' problem, i.e., the random division of the unit interval, and gives an asymptotic formula for the slope of the log-linear rank-size law in the upper tail of the log-normal distribution.

[1] Georg K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, New York, 1949) 1st ed.
[2] Benoît B. Mandelbrot, in *Communication Theory*, edited by Willis Jackson (Butterworths, London, 1953).
[3] Benoît B. Mandelbrot, in *Structure of Language and its Mathematical Aspects*, Proceedings of the Symposium on Applied Math, edited by Roman Jacobson (AMS, New York, 1961), Vol. 12.
[4] I. Kanter and D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).
[5] Benoît B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1983), 3rd ed.
[6] Benoît B. Mandelbrot, IRE Trans. Inf. Theory **3**, 124 (1954).
[7] Richard Perline, Phys. Rev. E **54**, 220 (1996).
[8] Allan Gut, *Stopped Random Walks, Limit Theorems and Applications* (Springer, New York, 1987).