

Mutual Information Functions of Natural Language Texts

Wentian Li

*Center for Complex Systems Research, Physics Department, Beckman Institute,
University of Illinois, 405 North Mathews Avenue, Urbana, IL 61801;
Department of Physics, Columbia University, New York, NY 10027;
and Santa Fe Institute, Santa Fe, NM 87501 **

August 22, 1989, revised on November 7, 1989

Abstract. The mutual information function $M(d)$, which is a quantity used to detect correlations in symbolic sequences, is applied to natural language texts. For some English and German texts being analyzed, $M(d)$'s for both the letter sequences and letter-type sequences exhibit approximate inverse power law function at shorter distance with exponents close to 3. This decay of $M(d)$ is too fast to lead a $1/f$ power spectrum. Due to finite size effects, it is not conclusive as to whether the same inverse power law function extends beyond short distances. Also included are discussions on various topics concerning other scaling phenomena in formal and natural languages.

1. Introduction

It was observed more than a decade ago [1] that the power spectra of the audio signal, for both the loudness and the pitch, of music and speech are approximately $1/f$ noise, i.e., random signals with inverse power law power spectra $P(f) \approx 1/f^\alpha$, with α being close to 1. Many other signals, ranging from star luminosity [2] to highway traffic flows [3] also show the same $1/f$ noise. For a review article, see Ref. [4].

We use scores to represent music and texts to represent speech. Not all the sequences with the score symbols or the alphabets can be performed or be read to have $1/f$ -like power spectra. For example, a "monkey-typed" text [5, 6, 7] is random, and has a constant power spectrum (i.e., white noise) if it could be read as a speech. Then, how does $1/f$ noise in music and speech emerge from the structures of the scores and the texts? In this paper, we will examine this question by carrying out statistical analysis of correlations and structures in several natural language texts.

The first obstacle in identifying $1/f$ noise in natural language texts is that the power spectrum cannot be defined for symbolic sequences. The

*Current address.

“complimentary” function of the power spectrum, the correlation function, which is the inverse Fourier transformation of the power spectrum, cannot be defined on symbolic sequences either. We will instead use the mutual information function [8] here for the analysis of language texts.

The mathematical and statistical analysis of language texts is not new (see, for example, Ref. [9]). Nevertheless, most of the previous studies are centered around the occurrence frequency of the language units (e.g., words) and the distribution of their length (e.g., sentences). There are also studies on the entropy [10, 11, 12, 13] and the conditional transition probability from one letter to its neighboring letter [10, 14]. These probability-based quantities appear more frequently in the literature after the information theory became popular. Although the analysis presented in this paper depends on the concepts in information theory, the emphasis is on the correlation between two letters separated by *any distance*, rather than a distance of one site (the Markov process approximation of a language is used on the belief that two neighboring sites are correlated with certain probability) or no distance at all (entropy does not account for the correlation between symbols).

In the next section, the mutual information function is introduced; Section 3 contains the numerical results on the mutual information function of many language texts; Section 4 discusses whether quantities calculated from left to right can detect structures generated from top to bottom; Section 5 examines the correlation functions in context-free languages, one of the formal languages, and relates them to that of the natural languages; Section 6 discusses other scaling laws in natural languages, including Zipf’s law; and a conclusion is drawn in the end of the paper.

2. Inverse Power Law Correlation In Symbolic Sequences

In this section, we briefly review the quantity for detecting correlations in symbolic sequences: mutual information function $M(d)$. In order to make comparisons with the more frequently used function, the correlation function, we proceed in our definition in the following way: Consider a numerical sequence $\{x_i\}$ ($i = 1, 2, \dots, N$), where the site value is taken from state variables $\{a_\alpha\}$ ($\alpha = 1, 2, \dots, M$), if the joint probability of any site i to take value a_α and site $j=i+d$ to take a_β is $P_{\alpha\beta}(d) \equiv \overline{P(x_i = a_\alpha, x_j = a_\beta)}$ (the overhead bar represents an average over site i), the correlation function is:

$$\Gamma(d) = \sum_{\alpha\beta} a_\alpha a_\beta P_{\alpha\beta}(d) - \left(\sum_{\alpha} a_\alpha P_\alpha\right)^2. \quad (2.1)$$

Now consider symbolic sequences where the state variables $\{a_\alpha\}$ are not numbers. We cannot just remove the expression $a_\alpha a_\beta$, or a_α in Eq. (2.1) in order to avoid the problem of using symbols, because the summation will then lead to the trivial value 1. The alternative is to define the following function, borrowing from the concept of mutual information in information

theory [10]:

$$\begin{aligned} M(d)^{[1]} &= \sum_{\alpha\beta} P_{\alpha\beta}(d) \log_2 \frac{P_{\alpha\beta}(d)}{P_\alpha P_\beta} \\ &= 2H^{[1]} - H^{[2]}(d), \end{aligned} \quad (2.2)$$

where $P_{\alpha\beta}$ is as defined in Eq. (2.1), P_α is the probability (density) for symbol a_α , $H^{[1]}$ is the entropy for a single site and $H^{[2]}(d)$ is the entropy for the “joint block” which is patched from two sites separated by a distance d . The definition can be easily extended to the mutual information between two 2-site blocks ($M(d)^{[2]}$), 3-site blocks ($M(d)^{[3]}$), etc.

We use the name mutual information function to emphasize that the quantity is a function of the distance between sites, making it closer to the name correlation function. Beware that the same name is used sometimes to refer to something different, for example, that mutual information is a function of the probability distribution (section 5.2 in Ref. [15]). The reason mutual information is a good quantity for measuring correlation is because, for a random sequence which has no correlations, the joint probability $P_{\alpha\beta}(d)$ is simply equal to the product of P_α , and P_β , so $\log_2(P_{\alpha\beta}(d)/P_\alpha P_\beta) = 0$, or $M(d)$ is zero. The same conclusion can also be drawn from the second line of Eq. (2.2): the entropy for blocks composed of two uncorrelated sites ($H^{[2]}(d)$) is twice that for a single site ($H^{[1]}$), and the two terms in $M(d)$ cancel to zero.

For numerical sequences, we can ask the question of whether a power law correlation function implies a power law power spectrum, or a power law mutual information function. The first part (from correlation function to power spectrum) can be easily answered: a $1/d^\beta$ correlation function corresponds to a $1/f^{1-\beta}$ power spectrum if $0 < \beta < 1$. A power law correlation function will also lead to an approximate power law mutual information function. This can be argued as follows: a power law $\Gamma(d)$ indicates a power law $P_{\alpha\beta}(d)$. $P_{\alpha\beta}(d) \log(P_{\alpha\beta}(d)/P_\alpha P_\beta)$ is then a power law function times a linear term, and it is *approximately* a power law function. A careful examination of binary sequences shows that the above conclusion is true [8]. After keeping only the leading terms in a large distance approximation, the exponent for the power law function in $M(d)$ is twice that for $\Gamma(d)$ [8]. Appendix 1 includes the relationship among the exponents of these 3 functions.

The way to relate symbolic sequences with $1/f$ noise is first to see whether the $M(d)$ for the symbolic sequence is an inverse power law function. If this is so, we then extrapolate the exponent for $M(d)$ to that of the corresponding $\Gamma(d)$, and then to see whether that $\Gamma(d)$ can lead to a $1/f$ power spectrum.¹ For binary sequences, it means a mutual information function $M(d) \sim 1/d^{2(1-\beta)}$ ($0 < \beta < 1$) is required for a $1/f^\beta$ power spectrum. For more general cases, a quantitative relation between the exponents of $\Gamma(d)$ and $M(d)$ is not currently available. Nevertheless, empirical evidence shows

¹The corresponding sequences can thus be called *symbolic 1/f noise*.

that $M(d)$ always decays faster than $\Gamma(d)$. We will see in the next section that although natural language does have inverse power law correlation, the exponent is too large to ensure a $1/f$ noise.

3. Analysis of Letter Sequences And Letter-Type Sequences

One problem encountered while analyzing natural language texts is how to decide on the unit of the symbol sequence. At the lowest level, a language text is considered as a string of letters including punctuations and blank spaces. Choosing letters as the minimum units is simple and easy in term of computation, because one does not need an extra set of instructions to transform the text into symbolic sequence with higher level units. The drawback for using letter sequence is that it is unlikely this analysis will provide a true understanding of the text, because a single letter usually does not carry a full meaning.

What we aim to do, at this stage, is to discover features common to all language texts (at least for English texts) at the level of letters. One can make a comparison with the power spectra analysis of music and speech done by Voss and Clarke [1]. In their study, all music, regardless of whether it is classical music or jazz, shows the same $1/f$ noise. Consequently the power spectrum analysis is less useful for understanding of each individual piece of music. On the other hand, one can identity all music as belonging to the same “universality class.” And one can separate music from random noise or any other signals which are not considered to be “music,” by means of power spectra.

In order to obtain the best statistics, I also analyze the “letter-type” sequences, i.e., the sequences with four symbols according to whether the letter is a vowel, consonant, punctuation, or space. This sequence is a “coarse-graining” of the original letter sequence.

One can also choose other units instead of letters, for example, word-types or “lexical categories” (with symbols of noun, verb, adjective, adverb, article, etc.), words, morphemes, and syllables. For language systems such as Japanese or Chinese, we deal with character sequences. In the languages in some ancient culture, the unit of the language is the ideogram. The problem with using these units is that it becomes more difficult to get good statistics in filling the histograms for the occurrence of each unit. Actually, the best and most logical choice of the unit should be the phoneme, not only because it is tractable (the number of phonemes is not very large), but also because every language can be spoken and, as a result, be recorded in phoneme sequences.

As a simple example for the illustration purpose, I have analyzed J. F. Kennedy’s inaugural speech (the text contains 7391 letters). Figure 1 shows what the correlation function (the absolute value) would look like, in log-log scale, if letters are transformed into numbers randomly. In the plot, there are 10 different $\Gamma(d)$ curves from different random transformations (though blank space is always mapped to 0 and punctuation is always mapped to 1). Clearly, the correlation function is not an invariant function with respect to

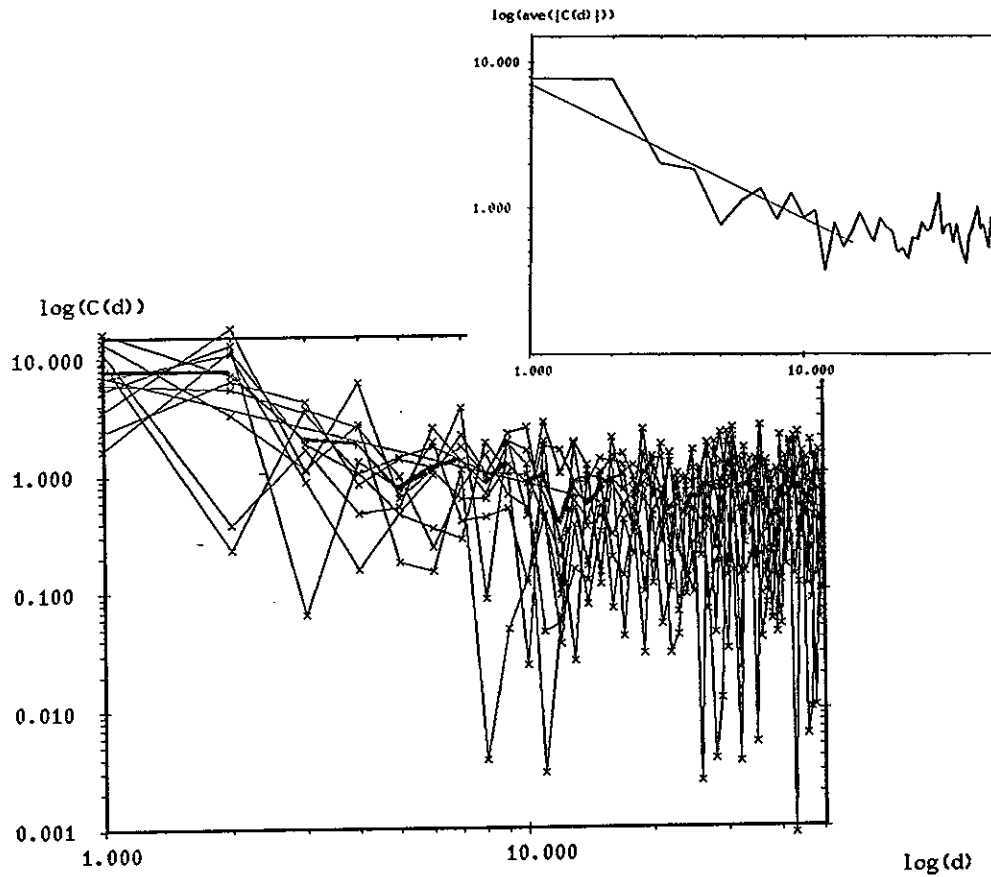


Figure 1: The absolute values of correlation functions for numerical sequences produced by randomly mapping 26 letters to numbers between 2 and 27, blank space to number 0 and punctuations, digits to number 1, from JFK's speech. The inset is the average of these ten functions, as well as the best fitting straight line for the first 15 points with slope -0.93 .

name	length	residue (4-1,4-2,28-1)
JFK's speech	7,391	0.002, 0.03, 0.1
Shakespeare's Hamlet	148,177	1×10^{-4} , 0.0017, 0.0053
AP's News	232,232	6.9×10^{-5} , 0.0011, 0.0034
Bible(German)	844,027	1.9×10^{-5} , 3×10^{-4} , 9.3×10^{-4}
Shakespeare's Plays	1,305,575	1.2×10^{-5} , 2×10^{-4} , 6×10^{-4}

Table 1: Source, the text length (number of letters) and the residue mutual information due to finite size effects.

the transformation. The inset is the average of all the $\Gamma(d)$'s, and the curve from $d = 1$ to $d = 15$ is fitted by a straight line with the slope -0.93 .

Figure 2 shows three $M(d)$ curves for the text of J. F. Kennedy's speech: site-to-site $M(d)^{[1]}$ for letter-type sequence (with 4 symbols), $M(d)^{[2]}$ for the same sequence, and $M(d)^{[1]}$ for the letter sequence (with 28 symbols). It is noticed that the mutual information functions do not decay to zero as expected. The existence of the non-zero residue values is because of the finite sequence length, which introduces fluctuations for counting the probabilities, and consequently an overestimation of the mutual information value [8]. As a crude estimation, the error of mutual information introduced by finite sequence length for purely random sequences is [8]:

$$\delta M(d) \sim + \frac{M^{2l}}{N}, \quad (3.1)$$

where N is the sequence length, M is the number of symbols, and l is the block length. Applying this formula to the three calculations of $M(d)$ on JFK's speech, we have the residues: $\delta M(d) \sim 0.002$ (letter-type sequence), 0.03 (letter-type, $M(d)^{[2]}$), and 0.1 (letter sequence). These residue values are drawn in the Figure 2. Also drawn are residue values calculated from the formula for 3-symbol and 27-symbol sequences for a comparison. Although the values calculated do not fit the data exactly because English texts are not random sequences (e.g., the 26 letters are not equally distributed), they do give a good estimation of the finite size effect.

We also notice that the $M(d)$'s decay as power laws at shorter distances. In order to test the validity of this observation, more natural language texts with much longer lengths are analyzed. They are: Shakespeare's play *Hamlet*, Associated Press news articles, a collection of 11 of Shakespeare's plays, and the Bible (in German). The information about the text lengths and the estimation of the residue of mutual information by Eq. (3.1) is included in Table 1. The information about the source and the editing procedures is included in Appendix 2.

Figure 3 shows $M(d)^{[1]}$ for the letter sequences taken from the four text files mentioned above, Figure 4 is the $M(d)^{[1]}$ for letter-type sequences from the same four text files, and Figure 5 is the $M(d)^{[2]}$ for the same letter-type sequences. We have the following comments:

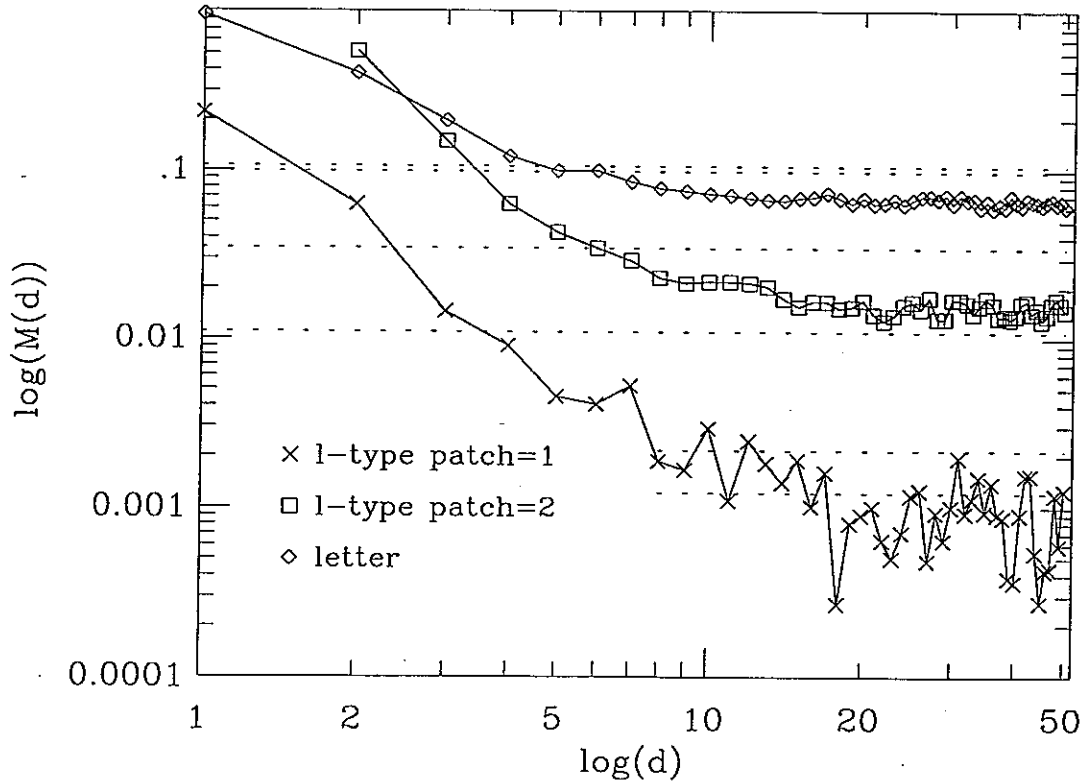


Figure 2: Mutual information function for the sequences transformed from JFK's speech: 1. letter-type sequence (with 4 symbols), $M(d)^{[1]}$; 2. letter-type sequence, $M(d)^{[2]}$; 3. letter sequence (with 28 symbols), $M(d)^{[1]}$. The dotted lines are estimated residue values of $M(d)$ for random sequences due to finite size effects (for both M -symbol and $(M - 1)$ -symbol).

- $M(d)$'s at shorter distances are similar for all texts (with minor variations for the German Bible). It might be explained by the fact that all the texts have the same or similar words structures. The curve at distance smaller than 5 letters deviates from the straight line at intermediate distances ($d \approx 5$ to 20) in the log-log plot, which means there might be two different functional forms for distances smaller or larger than the average length of English words (namely, 5). The exponents for the inverse power law function which approximates the $M(d)$ at intermediate distances are close to 3–3.5 for letter-type sequences. (The dashed lines in the plots all have slopes -3 .) It is still not firmly established what the exact value of the exponent is, because it fluctuates with different choices of fitting regions, different coarse-graining schemes, and different languages (the error bar for the exponent is roughly ± 1).
- The residue $\delta M(d)$ does indeed drop with the text length, as expected from Eq. (3.1). (The dotted lines in the plots are residue values for the corresponding random sequences.) In principle, by increasing the text length, one can gradually reach the edge of the scaling regions, but in practice, because the $M(d)$ decays very quickly to the residue value by a function of form $\sim 1/d^3$, increasing the text length by a factor of 10 can reduce the residue value by 10 and can only expand the distance by a factor of $\sqrt[3]{10} \approx 2.2$. Even for a very large language corpus, such as the Brown University Corpus of American English [16] with one million words or approximately 5 million letters,² the improvement on reducing the residue mutual information with respect to Shakespeare's plays used in this paper (one million letters) is only a factor of 5.

4. Hierarchical Structures In Natural Languages

This section is about the hierarchical structures in language texts and whether the mutual information function can detect such structures, especially whether an inverse power law $M(d)$ is related to such hierarchical structures. These discussions are mainly meant to raise the question, and motivate further investigations.

It is known that natural language texts are typical examples of hierarchical structure: letters are grouped into words; words are connected into sentences; sentences are organized as paragraphs, sections, chapters; and so on. Each level of the hierarchy has its own organization principle.

Although few rules for the construction of words are established, the probabilities of one letter to be followed by another letter is well studied ("first-order monkey" [6]). For example, the letter q should always be followed by the letter u in English. This Markov process or finite automaton approach is a reasonably good approximation for the structure of words.

²There are rumors, however, of the existence of a much bigger database [17].

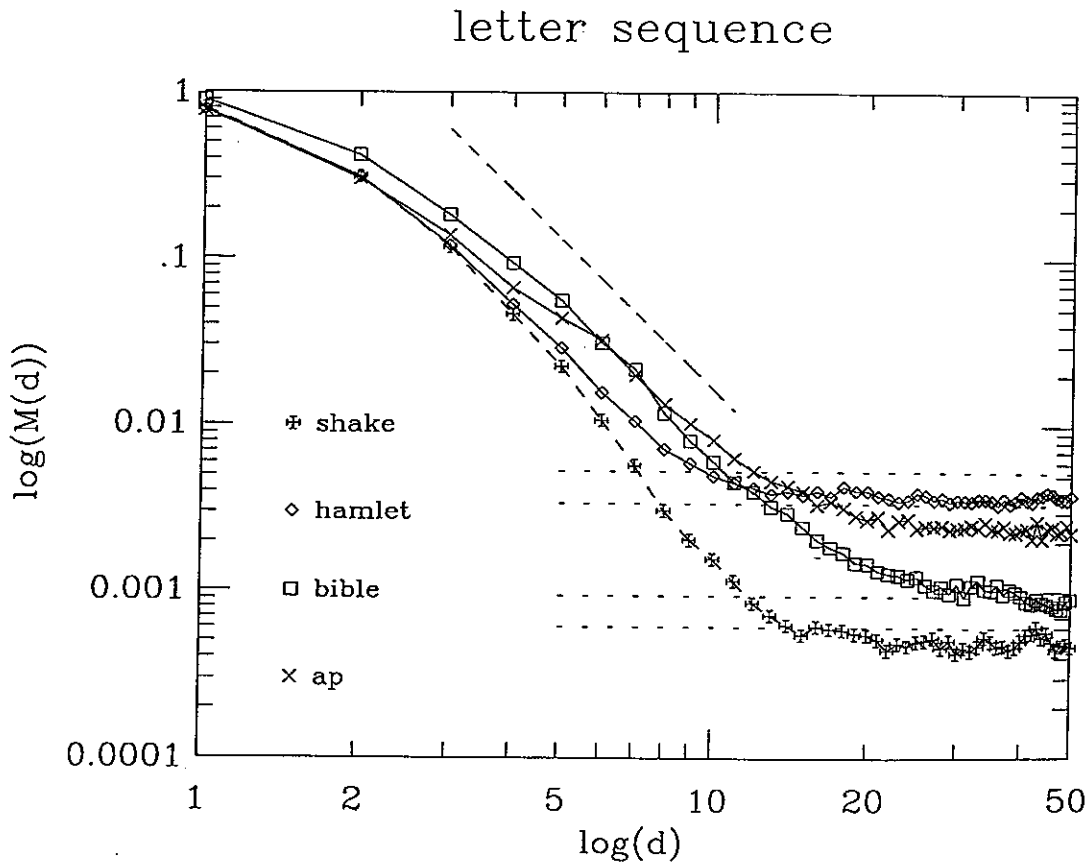


Figure 3: $M(d)^{[1]}$ for letter sequences from: 1. Shakespeare's play *Hamlet*; 2. Associated Press News Articles; 3. Five books of Moser from Bible (in German); 4. Eleven plays by Shakespeare. The dashed line is an inverse power law function $1/d^3$. The dotted lines are the estimated residue values of $M(d)$ for random sequences with the same sequence lengths and the same number of symbols as those of the texts.

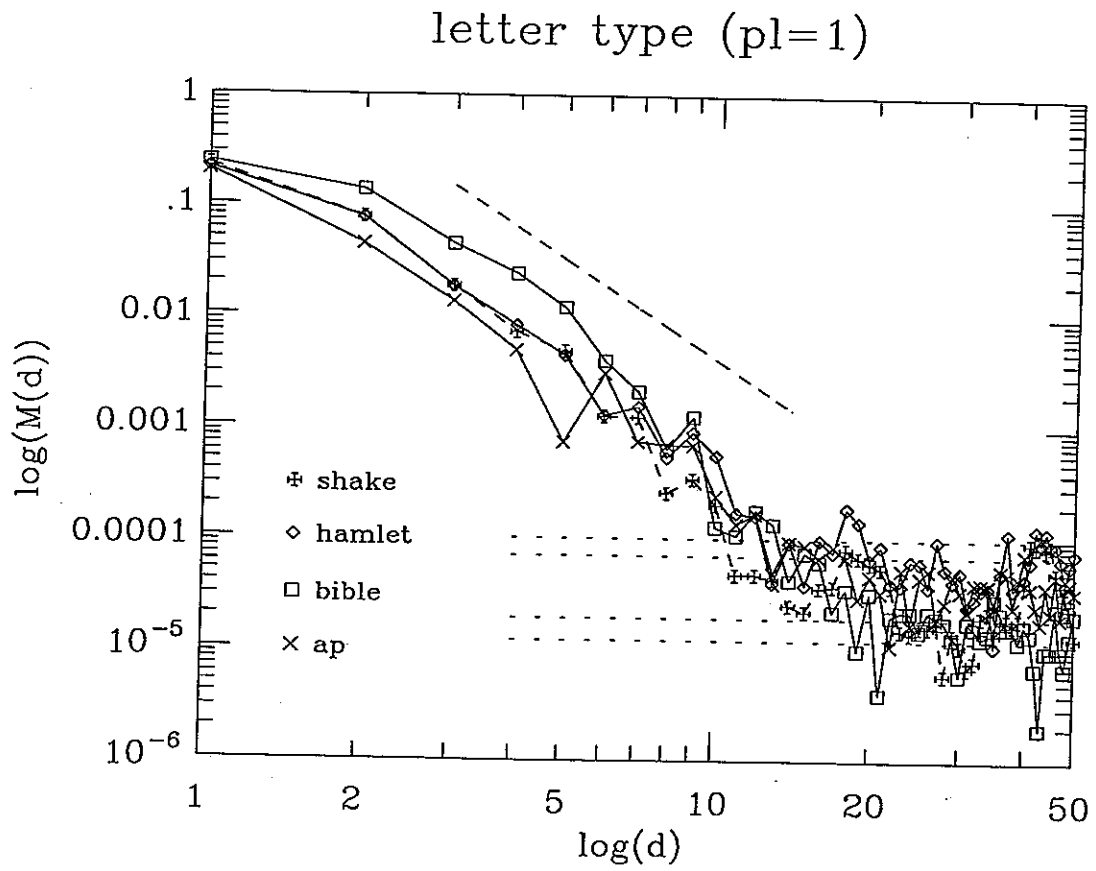


Figure 4: $M(d)^{[1]}$ for letter-type sequences. Same source as that in Figure 3.

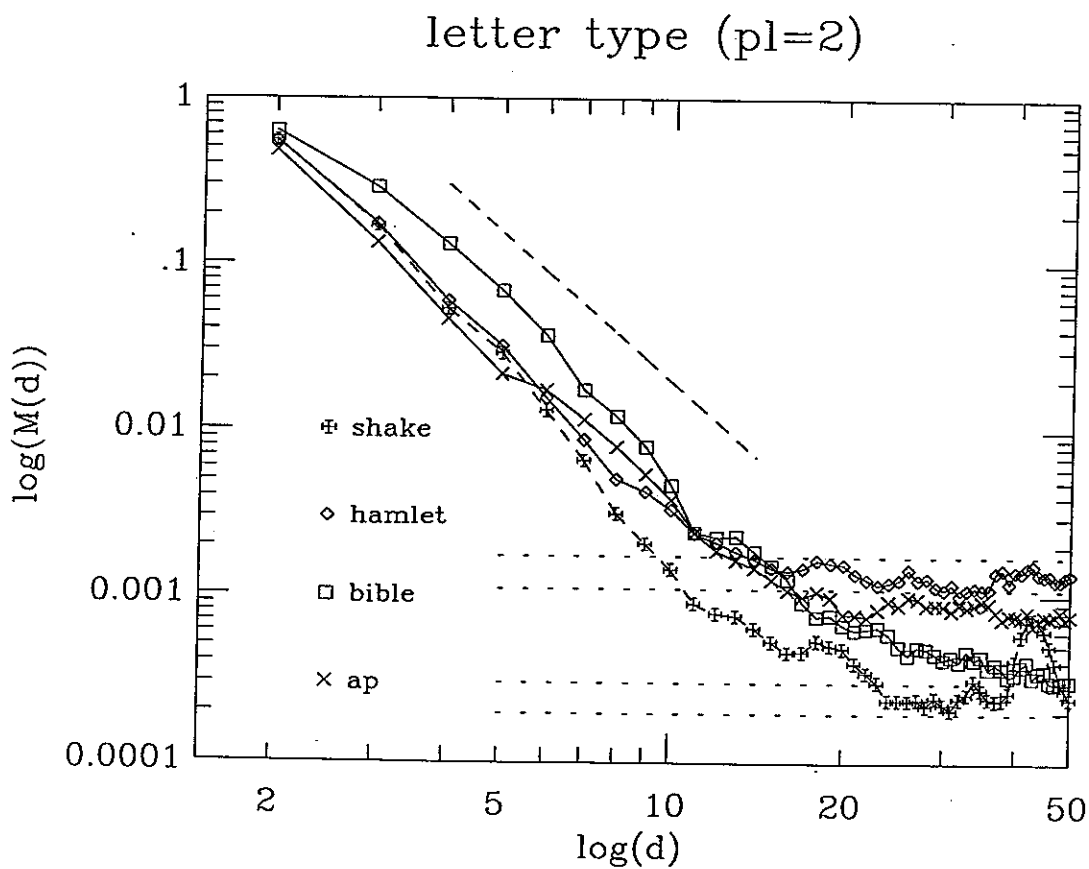


Figure 5: $M(d)^{[2]}$ for letter-type sequences. Same source as that in Figure 3 and 4.

The construction of sentences from words follows the phrase structure rules, which can be represented by a set of context-free grammar rules [18]. Although it has not been settled as to whether natural language grammar is indeed context-free [19], it is generally believed that context-free language is a good approximation of the natural languages. That the language units are organized in a hierarchical manner is most apparent from the fact that words are organized into sentences.

Above the level of sentences, the organization principle is more author-dependent; this is where many “style analysis” techniques come in [20]. For example, the average sentence length varies greatly from author to author. This point is nicely summarized by Herdan (p.117 of Ref. [9]): “Roughly we may say that the larger the linguistic unit which is chosen for the investigation, the more will the frequencies depend upon individual style.” We do not expect any universal result.

For levels above sentences and paragraphs, it is unlikely that a pure sequence analysis will measure correlations appropriately. The relationship between different parts of a language text is largely determined by the semantics (the meaning of the words, phrases, and sentences), and pragmatics [21] (the correct usage of the languages). In order to understand the meaning of words, one needs something else, for example, the dictionary. Besides, there are “hidden” connections which are set up by the logic developed in the text. For example, the last page of a detective story will usually reveal the relationship among characters and facts which is obscure on the superficial level. Obviously, the mutual information function cannot detect such relationships.

From the above discussions on the hierarchical structures of natural languages, one conclusion is that $M(d)$ may have different functional forms at different distances because the corresponding organization principles are different. It would certainly be nice for the inverse power laws of $M(d)$ to have the same exponent at all distances, but there is no reason to believe such is the case.

When the language unit we choose is too small (e.g., letters), it is not clear whether the very large-scale structures in language can be detected. It is analogous to the situation of the human body, which is not best analyzed on the smallest level such as atomic or molecular level. In order to cross the gaps between different levels, one can increase the size of the unit and calculate the mutual information between these larger units, though in practice the statistics become worse since we have less number of countings to fill up the histogram.

In summary, hierarchical structures are much easier to detect “from top to bottom” simply because it is how they are generated. The mutual information function of a sequence is calculated from left to right, and thus it is somewhat limited in its ability to detect hierarchical structures. What I have suggested to compensate this is to change the size of the unit, and calculate the mutual information for the new units. Finally, it should be remembered that it is unlikely to use a simple numerical function such as $M(d)$ to detect

correlation between meanings in language texts. It should not be surprising that we fail to detect long-range correlations expected by the logical connections between sections of an article by measuring $M(d)$. For the same reason, even if computers were equipped with the correct syntax rules, they still could not always produce appropriate messages.

5. Correlations In Context-Free Languages

In this section, we will discuss correlations in context-free languages with certain grammar. Unlike regular languages and Markov processes, which produce sequences with exponentially decaying correlation functions [22], it is typical for context-free languages to produce sequences with inverse power law correlation functions. One example is a particular context-free language,³ called the expansion-modification system, which can generate sequences with $1/f^\alpha$ ($\alpha \approx 1$) power spectra and inverse power law correlation functions [24].

The prototype rule for the expansion-modification system is defined as the following: $0 \rightarrow 00$; $0 \rightarrow 1$; $1 \rightarrow 11$; $1 \rightarrow 0$, with probabilities $1 - p, p, 1 - q, q$ respectively [24]. The tuning of the parameters p, q can produce sequences with completely different features. In two extreme cases: (i) $p = q \approx 0$, the expansion factor dominates the dynamics, and the sequences produced tend to have “structures” with very long length (here the structures are simply the strings of 0’s or strings of 1’s); (ii) $p, q \approx 1$, the modification part of the dynamics dominates, and the resulting sequences are random. In between, when p, q are small but non-zero (e.g., $p = q = 0.1$), the “structures” with longer scales can either be generated by expansion or destroyed due to the modification, which splits the long string of 0’s (or string of 1’s) from inside. When the sequence is examined at a fixed time step, its power spectrum scales perfectly as a straight line in log-log scale with slope very close to -1 .

The basic constituent structure for English sentences can be generated by context-free languages such as $S \rightarrow NV$ (sentence is composed of a noun phrase and a verb phrase), $N \rightarrow S$ (a noun phrase can be a sentence itself), $N \rightarrow n$ (a noun phrase can be a noun), $V \rightarrow vN$ (verb followed by a noun phrase), $V \rightarrow v$ (a verb phrase can be a verb), and so on.

The grammar rules for natural languages are usually iterated only a few times, unlike the typical study of “attractors” of expansion-modification systems and other dynamical systems, in which sequences are produced by iterating until the time-invariant states are reached. If the grammar rules are applied too many times, the sentences will be too complicated for normal people to comprehend. Another difference between natural language grammar and the expansion-modification system is that the leaves of the natural language grammar tree can stop at different time steps for different branches, whereas those for the latter all stop at the same time. One can make the two similar by adding extra rules in the natural language grammar to preserve

³More accurately, it is a context-free Lindenmayer system, which does not distinguish terminal symbols from non-terminal symbols, and which updates all sites simultaneously [23].

the non-terminal symbols along the branch (e.g., $v \rightarrow v$).

Other than these two differences, the natural language grammar is weakly reminiscent of the expansion-modification systems. In particular, the $S \rightarrow NV$ is similar to expansion, and the $N \rightarrow S$ is similar to modification. Whether the detailed form of $M(d)$ for word sequences can be formally established by a mathematical study of natural language grammar rules is still under investigation. Even if the $M(d)$ can be determined for word sequences, it is still not clear how to transform it to that of the letter sequences.

6. Other Power Law Scalings In Natural Languages

The main purpose of this paper is to find some power law relation in natural language texts in order to characterize its hierarchical structures (the mutual information function is one of the examples), so we might examine some other possible power laws. We will discuss in this section three such possibilities: (1) concepts related to fractals, (2) block entropy as the function of block length, and (3) Zipf's scaling law.

Fractals as applied to geometric objects are related to the scaling relation between the number of non-empty boxes and the box size. The fractal dimension is the exponent for this scaling relation. For symbolic sequences such as language texts, this definition cannot be applied without modifications. One alternative is to partition the sequence into non-overlapping blocks and look at how the number of possible block configurations $N(n)$ changes with the partition width. Whether $N(n) \sim e^{\alpha n}$, or $\sim e^{n^\alpha}$ ($0 < \alpha < 1$, stretched exponential), or $\sim n^\alpha$, we can always extract some exponent α as an analogy to the fractal dimension.

In the case of overlapping blocks, we actually deal with the topological block entropy $S(n) = \log N(n)$ (or metric entropy $S(n) = -\sum p_i^{[n]} \log p_i^{[n]}$, with $\{p_i^{[n]}\}$ are probabilities for all configuration of n -blocks) as a function of the block length. Again there are several possibilities: $S(n) \sim \alpha n$, or $\sim n^\alpha$, or $\sim \alpha \log(n)$. The first case is for random sequences (although it does not have to be purely random), and the second and third cases are somehow between random and regular. The second type of divergence of block entropy has been observed for sequences generated by cellular automata [25] and by expansion-modification systems [26]. In general, the block entropy may not be a simple function of the block length. For example, it might be a mixture of power law and logarithm function $n^\alpha \log(n)^\beta$ [27]. It should be noted that block entropies are derived from the “ n -point correlation” (n -site joint probability distribution), whereas the site-to-site mutual information function is a “2-point correlation” measure. In a sense, the block entropy contains more information about the sequence than the site-to-site mutual information function.

The last example of power law scaling in languages is of a quite different nature: it is the scaling between the occurrence frequency of words and the ranks (e.g., the most frequently occurring word ranks first, the second most frequent one ranks second, and so on) of the word. It is claimed by Zipf

that this relation between the frequency and the rank should be an inverse power law with exponent equal to 1 [28]. The claim is checked to be of very good accuracy by database analysis (e.g., Ref. [29]). Unfortunately, what Zipf has overlooked is that even monkey-typed random texts obey Zipf's law [30]. A simple proof will be given in Appendix 3. So, Zipf's law is not some feature which is related to the hierarchical structures of natural languages.⁴ Consequently, it has nothing to do with the inverse power law for mutual information functions.

Conclusion

To conclude, I calculate the mutual information function for letter sequences and letter-type sequences of some English texts (and German texts, too), which behave as inverse power law functions with exponents around 3. The mutual information function with this exponent will not lead to $1/f$ noise. Further studies should be done on the mutual information function for more meaningful sequences transformed from language texts, for example, a phoneme sequence. The strikingly similar curves of $M(d)$ for different text files indicate a universality class of natural language, which distinguishes them from random letter sequences and other sequences with short-range correlations. It is reminiscent of another "universality class" of music and audio signal of speech which is characterized by the appearance of $1/f$ noise.

Acknowledgement

I appreciate my discussion with the following people about the linguistics issues: Howard Macley, C. C. Cheng, and Jerry Morgan. I would thank Fred Richards, Atlee Jackson and Chris Langton for proofreading an earlier draft. I thank Christoph Schulz for transfer the German Bible to me, Dr. I. Tsuda for sending me his paper before publication, Salmon Burkie for discussions on the Zipf's law for random texts, and Fred Richards for many valuable suggestions. The work is supported by National Science Foundation Grant PHY-86-58062, and Office of Naval Research Grant N00014-88-K-0293. Part of the work is done in Santa Fe Institute, and I acknowledge the support from NSF Grant PHY-87-14918 and DOE Grant DE-FG05-88ER25054.

Appendix 1: Relation Between Exponents

In this section, we will show that if one of the three functions— power spectrum, correlation function, and mutual information—is an inverse power law with the exponent in a certain range, the remaining two are also inverse power law functions. More specifically, if the power spectrum is $P(f) \sim 1/f^{1-\beta}$ ($0 < \beta < 1$), the correlation function will be $\Gamma(d) \sim 1/d^\beta$ (true for any sequences), and the mutual information function is $M(d) \sim 1/d^{2\beta}$ (true for binary sequences).

⁴There are some misunderstanding about this in the literatures, e.g., Ref. [31].

The definition of power spectrum $P(f)$ is the Fourier transformation of the correlation function:

$$P(f) = \frac{1}{N} \sum_{d=0}^{N-1} e^{-i2\pi(f/N)d} \Gamma(d). \quad (6.1)$$

Suppose the correlation function is $\Gamma(d) \approx 1/d^\beta$, which satisfies $\Gamma(d) \sim k^\beta \Gamma(kd)$, we have:

$$\begin{aligned} P(f) &= \frac{1}{N} \sum_{d=0}^{(N-1)/k} e^{-i2\pi(f/N)d} k^\beta \Gamma(kd) \\ &= \frac{k^\beta}{k} \frac{1}{(N/k)} \sum_{d'=0}^{(N-1)/k} e^{-i2\pi(f/kN)d'} \Gamma(d') \approx k^{-(1-\beta)} P(k^{-1}f) \end{aligned} \quad (6.2)$$

($d' = kd$). It is a power law $P(f) \approx 1/f^{1-\beta}$ as claimed above.

The second relation among exponents is only proved for binary sequences. Suppose the binary sequence has state value 0 and 1. The correlation function has the form $\Gamma(d) = P_{11}(d) - P_1^2 \approx 1/d^\beta$. It can be proved that other joint probabilities satisfy [8] $P_{00} = \Gamma(d) + P_0^2$, and $P_{01} = P_{10} = -\Gamma(d) + P_0 P_1$. Replace these $P_{\alpha\beta}$ in the mutual information function and take the large distance limit, we have [8]:

$$M(d) \approx \frac{1}{2} \left(\frac{\Gamma(d)}{P_0 P_1} \right)^2 \approx \frac{1}{d^{2\beta}}. \quad (6.3)$$

Appendix 2: Preparation of the Texts

This appendix includes the source of the texts and the editings that I did for them. In all cases, capital letters are switched to lowercase letters, and all digits are considered to be punctuations.

- **J. F. Kennedy's Inaugural Speech:** The source is *Great American Speeches: 1898-1963*, edited by John Graham (ACC, 1970). I inserted extra symbol zeros between every two paragraphs.
- **Associated Press News Articles:** This is a collection of a dozen news articles about China's events from June 3 to June 9, 1989 by John Pomfret, Terril Jones, Jim Abrams, Dan Bires, Christopher Connell, Larry Thorsom, Elaine Kurtenbach, Susan Ruel, Kathy Wilhelm, and Howard Goldberg. There is no special arrangement among these articles. I simply delete all the titles, authors, time of issue, etc., and leave no space between different articles. Other minor editing has also been carried out, for example, changing all double quotation marks to single quotation marks; deleting all commas that separate digits (e.g., 5,000 is changed to 5000), etc.

- **Shakespeare’s Plays:** The source is from *William Shakespeare: The Complete Works*, edited by Stanley Wells and Gary Taylor (Oxford University Press, 1986). The digital edition is prepared by Catherine Smith and Michael Hawley (NeXT Digital Press, 1988). I deleted all the players’ names at the beginning of every dialogue, as well as the scene descriptions (e.g., “The Queen falls down”). The collection includes 11 plays: *All Is True*, *Hamlet*, *Henry V*, *Macbeth*, *Richard II*, *Richard III*, *Romeo and Juliet*, *The Merchant of Venice*, *The Taming of the Shrew*, *The Tragedy of King Lear*, and *Twelfth Night*.
- **Bible (in German):** The text includes the 5 books of Moses from Elberfeld Translation, which originally appeared in 1871. The revised version appeared in 1985, and the digital edition is from R. Brockhaus Verlag. I deleted all the section and chapter numbers. All the umlauts (the two dots on top of vowels) are deleted (e.g., ü becomes *u*). Also all the letter ß’s are deleted.

Appendix 3: A Proof That Random Texts Obey Zipf’s Law

This appendix will show that randomly generated texts obey Zipf’s law [28], i.e., the occurrence of a word is inversely proportional to its rank. Suppose there are M symbols, and one of the symbols is labeled as “blank space” ($_$). A word is defined as the symbol string between any two blank spaces (e.g., “_qwiu_”).

For randomly generated strings, any symbol occurs with equal probability $= 1/M$. The probability for words with length n to appear is equal to the probability that two blank spaces are separated by the distance $n + 1$: $P(n) \sim (1/M)^{n+2}$, or $P(n) = C/M^n$ (C is a constant). It shows that words with shorter lengths occur more frequently, i.e., they have higher ranks.

There are M^j words with length j . Because those words with shorter lengths have higher ranks, the ranks of words with length n ($i(n)$) satisfy

$$\sum_{j=1}^{n-1} M^j < i(n) \leq \sum_{j=1}^n M^j \quad (6.1)$$

or,

$$\frac{M}{M-1}(M^{n-1} - 1) < i(n) \leq \frac{M}{M-1}(M^n - 1). \quad (6.2)$$

It can be converted to

$$n - 1 < \log_M \left(\frac{M-1}{M} i(n) + 1 \right) \leq n. \quad (6.3)$$

Inserting this to the formula of probability of words with length n , we have:

$$P(i(n)) \leq \frac{CM}{M-1} \left(\frac{1}{i(n) + M/(M-1)} \right) < P(i(n-1)) \quad (6.4)$$

which is a “generalized” inverse power law distribution. Since $M/M-1 = 28/27 \approx 1$, this generalized inverse power law is a good approximation of the power law when $i(n) \gg 1$.

References

- [1] R. Voss and J. Clarke, "1/f Noise in Music and Speech," *Nature* **258**, 317-318 (1975).
- [2] P. L. Nolan, et al. "Rapid variability of 10-140 keV X-rays from Cygnus X-1," *The Astrophysical Journal* **246**, 494-501 (1981).
- [3] T. Musha and H. Higuchi, "The 1/f fluctuation of a traffic current on an expressway," *Jap. J. Appl. Phys.* **15**(7), 1271-1275 (1976).
- [4] E.g., William H. Press, "Flicker Noise in Astronomy and Elsewhere," *Comments On Astronomy* **7**(4), 103-119 (1978).
- [5] Sir Arthur Eddington, "The Nature of the Physical World," The Gifford Lectures, Cambridge (1927).
- [6] William Ralph Bennett, Jr. *Scientific and Engineering Problem-solving with the Computer* (Prentice Hall, 1976).
- [7] Brian Hayes, "A Progress Report on the Fine Art of Tuning Literature into Drivel," Computer Recreations, *Scientific American* **249**(5), 18-28 (1983).
- [8] Wentian Li, "Mutual information functions versus correlation functions" (Technical Report 89-1, Center for Complex Systems Research, Univ. of Illinois, 1989, submitted).
- [9] Gustav Herdan, *Type-Token Mathematics : A Textbook of Mathematical Linguistics* (Mouton, 1960).
- [10] C. E. Shannon, "The Mathematical Theory of Communication," *Bell Syst. Techn. Journal* **27**, 379-423 (1948).
- [11] C. E. Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. Journal* **50-64** (1951).
- [12] Thomas M. Cover and Roger C. King, "A Convergent Gambling Estimate of the Entropy of English," *IEEE Transactions on Information Theory* **IT-24** (4) 413-421 (1978).
- [13] Peter Grassberger, "Estimating the information content of symbol sequences and efficient codes," (Univ. of Wuppertal preprint, WU-B-87-11, 1987).
- [14] A. A. Markov, "Primer statističeskogo issledovanija nad tekstom 'Evgenija Onegina' illjustrirujuščij svjaz' 'ispytanij v cep'," *Izvestija Imper. Akademii nauk*, series VI, y. VII (3) (1913).
- [15] Richard E. Blahut, *Principles and Practice of Information Theory* (Addison-Wesley, 1987).
- [16] W. N. Francis and H. Kučera, *Manual of information to accompany a standard corpus of present-day edited American English for use with digital computers* (Department of Linguistics, Brown University, 1964, revised 1971 and 1979).

- [17] Roger Garside, Geoffrey Leech, and Geoffrey Sampson, eds., *The Computational Analysis of English: A Corpus-based Approach* (Longman, 1987).
- [18] N. Chomsky, "Three Models for the Description of Language," *IRE Transactions on Information Theory IT-2*, 113-123 (1956).
- [19] Geoffrey K. Pullum and Gerald Gazdar, "Natural Languages and Context-free Languages," *Linguistics and Philosophy* 4, 471-504 (1982).
- [20] Lubomír Doležel and Richard W. Bailey, eds., *Statistics and Style* (Elsevier, 1969).
- [21] E.g., S. Levinson, *Pragmatics* (Cambridge University Press, 1983).
- [22] Wentian Li, "Power spectra of regular languages and cellular automata," *Complex Systems* 1 107-130 (1987).
- [23] E.g., Arto Salomaa, *Formal Languages* (Academic Press, 1973).
- [24] Wentian Li, "Spatial $1/f$ Spectra in Open Dynamical Systems," to be published in *Europhysics Letters* (1989); and "Context-free Languages Can Give $1/f$ Spectra," (Technical Report 88-9, Center for Complex Systems Research, University of Illinois, 1988).
- [25] Peter Grassberger, "Long-range Effects in an Elementary Cellular Automaton," *Journal of Statistical Physics* 45 (1/2), 27-39 (1986).
- [26] Wentian Li and Tom Meyer, unpublished (1988).
- [27] P. Gaspard and X. J. Wang, "Sporadicity: Between periodic and chaotic dynamical behaviors," *Proceedings of National Academy Science USA* 85, 4591-4595 (1988).
- [28] G. Zipf, *Selective Studies and the Principle of Relative Frequency in Language* (Cambridge, Mass, 1932); *Human Behavior and the Principle of Least-Effort* (Cambridge, Mass, 1949; Addison-Wesley, 1965); *The Psycho-biology of Language: An Introduction to Dynamic Philology* (Houghton Mifflin Company, 1935; MIT Press, 1965).
- [29] Henry Kučera and W. Nelson Francis, *Computational Analysis of Present-Day American English* (Brown University, 1967).
- [30] George Miller, Introduction in *The Psycho-biology of Language: An Introduction to Dynamic Philology* (MIT Press, 1965).
- [31] John S. Nicolis and Ichiro Tsuda, "On the parallel between Zipf's law and $1/f$ processes in chaotic systems possessing co-existing attractors — A possible mechanism for language formation in the cerebral cortex," *Progress in Theoretical Physics* 82(2) (1989).