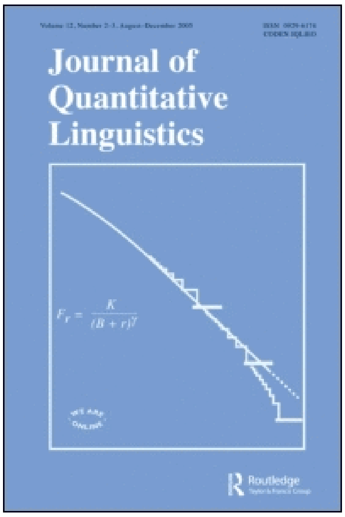


Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100702

Analysing the Hierarchical Organization of Text by Using Biologically-Inspired Statistical Methods

Israela Becker [a]; Eli Flaxer [ab]
[a] AFEKA - Tel-Aviv Academic College of Engineering, Tel-Aviv, Israel [b] School of Chemistry, The Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv, Israel

Online Publication Date: 01 November 2008

To cite this Article Becker, Israela and Flaxer, Eli(2008)'Analysing the Hierarchical Organization of Text by Using Biologically-Inspired Statistical Methods',Journal of Quantitative Linguistics,15:4,318 — 339

To link to this Article: DOI: 10.1080/09296170802326657

URL: http://dx.doi.org/10.1080/09296170802326657

Journal of Quantitative Linguistics
2008, Volume 15, Number 4, pp. 318–339
DOI: 10.1080/09296170802326657

Routledge
Taylor & Francis Group

Analysing the Hierarchical Organization of Text by Using Biologically-Inspired Statistical Methods*

Israela Becker[1] and Eli Flaxer[1,2]
[1]AFEKA - Tel-Aviv Academic College of Engineering, Tel-Aviv, Israel; [2]School of Chemistry, The Raymond and Beverly Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv, Israel

ABSTRACT

We have binarily encoded written monologues in a similar manner to the encoding of a continuous electrical signal of neuronal networks. The encoding keys have been carefully designed to include phrases that are co-referents. Only such keys produce binary series, which have an hierarchical fractal structure. This fractal structure can be revealed by using a statistical tool named "Fano factor". By outlining the analogy between the hierarchical structure of neuronal electrical activity and the hierarchical structure of text, we try to link the organization of text and neuronal activity.

INTRODUCTION

One of the classic phenomena in natural sciences is the existence of the ubiquitous "1/f" behaviour in many natural systems. The "1/f" terminology arises from an examination of the statistical properties of temporal fluctuations of many systems, whose activity is revealed on many time scales. Such temporal behaviour is analysed in the frequency domain rather than in the time domain. A low frequency is the equivalent of a long time scale, and a high frequency is the equivalent of a short time scale. Accordingly, if one draws the power spectrum of a system, which reveals behaviour on many time scales, one encounters 1/f behaviour.

*Address correspondence to: Israela Becker, AFEKA, Tel-Aviv Academic College of Engineering, 218 Bney-Efraim Rd., 69107 Tel-Aviv, Israel. Tel: 972-3-7688726.
Fax: 972-3-153-6994915. E-mail: IsraelaB@afeka.ac.il

Specifically, the probability of the occurrence of long time scales (i.e. low frequencies) *does not strongly decay* to zero.

It has long been suggested (Bak et al., 1987, 1988) that systems dominated by low frequencies consist of many interacting components. These systems, under many conditions, organize themselves into a state, which is complex (yet governed by local and simple rules) and has a rather general structure over many time scales. Such general structure indicates that the temporal fluctuations appear in the same proportion on all time scales. In this sense, these systems are considered self-similar entities, in which long-duration temporal correlation appears. Such self-similar structures are commonly referred to as fractal systems.

Neuronal electrical activity has been recorded and analysed in many experiments. The results of such analyses suggest the existence of long-duration temporal correlations along the recorded data. Therefore, these systems are regarded as fractal systems. As such analyses, being statistical, are performed on the entire set of the collected electrical data, and the results of such analyses show long-duration neuronal electrical activity, the entire set of collected data is considered a single body of information.

It has recently been observed that the temporal analysis of the oculomotor motion, while reading a text, follows a fractal structure (Schmeisser et al., 2001). The act of writing, or more specifically, the spatial distribution of intervals between signs in hand-written Japanese texts, has also been proven to show a fractal behaviour (Saiki et al., 1999). Text, like the phenomena of reading and writing, is brain-driven. As such, one expects to see a similar fractal structure appearing in a text.

The methods used to reveal the fractal nature of the electrical activity of neuronal networks and the acts of reading and writing can be used to examine the characteristics of text. These analyses consider the electrical activity of a neuronal network as a data set, which is a single body of information. Hence, they may be used equally on text units, which are defined as sequences of sentences that are intentionally logically and rhetorically related to form a single body of information.

In order to present this form of analysis and discuss its implications with regard to text, we first review the experimental work and data-set construction performed on electrical activity signals of neuronal networks, where the activity of neuronal networks is encoded to form binary strings. Secondly, we describe the statistical method used to analyse the binary string of the electrical activity of neuronal networks and its physical interpretation. Thirdly, we briefly discuss the concept of anaphora that is a pivot in text analysis. The anaphora will enable us to binarily display the chosen texts, similarly to the binary representations of neuronal networks' electrical activity. Finally, we apply the statistical method, described earlier, on a characteristic literary text, and discuss the results of the analysis and their implications.

RELATED EXPERIMENTAL WORK AND DATA-SET CONSTRUCTION OF NEURONAL SPIKE TRAINS

One of the motivations for the statistical analyses of the electrical data recorded from neuronal networks was the attempt to provide an answer to the intriguing issue as to where information is hidden in the electrical activity of neurons. For the purpose of analysis, the analogue-continuous form of the recorded neuronal membrane voltage was encoded into a binary sequence of idealized impulses (spikes). In such a binary sequence the 1s represent electrical signals that are higher than a certain threshold, and the 0s electrical signals lower than that threshold. Thus, the analogue signal is reduced into a point process. Figure 1 (a, b) is a diagram of the binary encoding procedure.

Two sequences can then be constructed out of the resulting idealized spike train:

- A sequence of number of counts (count train), that is constructed by specifying the number of spikes that occur in successive counting periods, each of duration *T*.
- A sequence of inter-spike times (inter-spike interval train) that is constructed by specifying the length of the time intervals between successive spikes.

Figure 1 (c, d) is a diagram of the construction of data sequences out of the encoded binary sequence of idealized impulses (spikes).

Both spike trains are then statistically analysed. The count sequence is used to measure the electrical firing rate and the spike-number histogram. The inter-spike interval sequence is used to calculate the inter-event interval (IEI) histogram (Cox & Lewis, 1996). The latter is also used to demonstrate the procession of events. Based on these two spike trains, researchers have re-defined the intriguing question as to where the

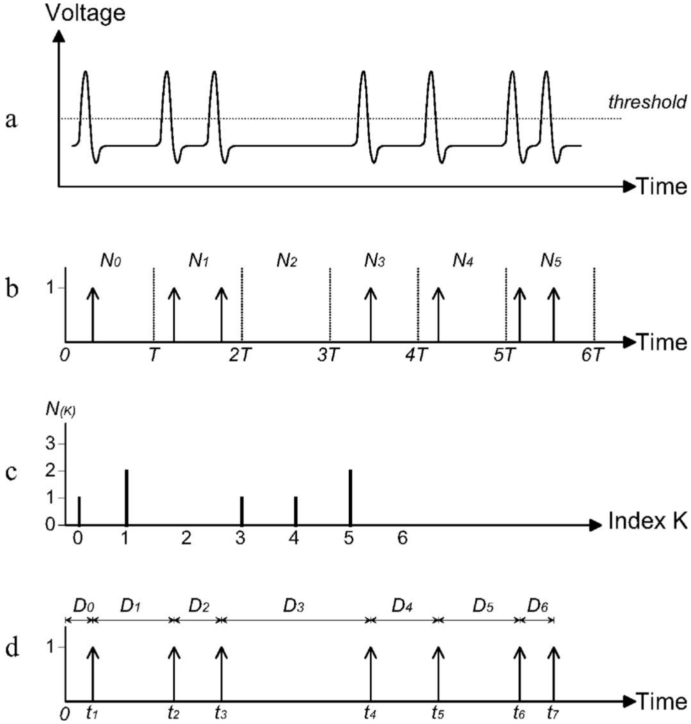


Fig. 1. The data-set construction of a binary data sequence out of the analogue-continuous form of the recorded neuronal membrane voltage. The continuous membrane voltage waveform (a) is idealized to a point process in which the occurrence times of action potentials are represented as discrete events (impulses), symbolized by arrows, on the time axis (b). The time axis can be divided into fixed intervals of width T, in which the number of events is counted to form a sequence of event counts (N_0, N_1, N_2, N_3, ... , N_n) in each time interval (c). A sequence of inter-event intervals (D_0, D_1, D_2, D_3, ... , D_n) is also constructed from this point process by specifying the width of each interval as a function of its index (d).

information is hidden within the electrical spike train: is it within the sequence and rate of spikes or is it within the sequence of intervals between the spikes?

Analyses of the inter-spike interval trains of the electrical activity of certain biological systems reveal the existence of long-duration temporal correlations (also referred to as "memory"; the term "memory" is written in commas as it has a different meaning from the term "memory" that is used in psychological research). Long-duration temporal correlations in such systems appear both at the microscopic (neurotransmitter exocytosis at the synapse; Lowen et al., 1997) and macroscopic (fluctuations in the sequence of human heartbeats; Turcott & Teich, 1993) levels. Extensive measurements of such an electrical activity have also been performed on animal sensory-system neurons such as cat retinal ganglion and lateral-geniculate (Teich et al., 1997), cat medullary sympathetic neurons (Lewis et al., 2001) and primary afferent auditory neurons of cat (Lowen & Teich, 1996), chinchilla (Powers et al., 1992) and chicken (Powers & Salvi, 1992).

Based on the results of analyses on a large corpus of data, it is now commonly accepted (Teich et al., 1996) that in a spike sequence of neuronal activity, which has a fractal structure, the occurrence times of the spikes carry all the information possessed by the spike train.

THE FANO FACTOR

A common tool to characterize such spike trains and a useful count-based measure of the autocorrelation of point processes is the Fano factor (Fano, 1947). This factor provides a direct estimation of the self-similarity of a given sequence. Since neuronal spike trains are regarded as stochastic point processes, they are frequently analysed using the Fano factor.

If one defines $N_i(T)$ as the number of events that occur in the ith time window of length T, the Fano factor, $F(T)$ is defined as the variance of $N_i(T)$ divided by the mean of $N_i(T)$, for a given window of length T:

$$F(T) = \frac{Var[N_i(T)]}{mean[N_i(T)]}$$

A curve is then constructed by plotting the logarithm of the Fano factor as a function of the window size.

Three kinds of processes can be distinctly discerned by using the Fano factor plot:

1. For a "renewal" point process, more often referred to as "random distribution", in which the inter-event intervals (IEIs) are non-correlated, the Fano factor is 1.0 for all window sizes.
2. For a "periodic" point process, the Fano factor approaches 0, due to the small value of variance of the number of events and a

simultaneous increase in the mean number of events, as the window size increases.

3. For a "fractal" point process, the Fano factor increases as a function of the window size. This increase reflects the greater variance of event counts with increasing window size due to the appearance of longer time scales (longer IEIs) as the window size increases.

Thus, the Fano factor serves as a sensitive indicator for the presence of a fractal structure. Figure 2 shows Fano factor calculation vs. the window

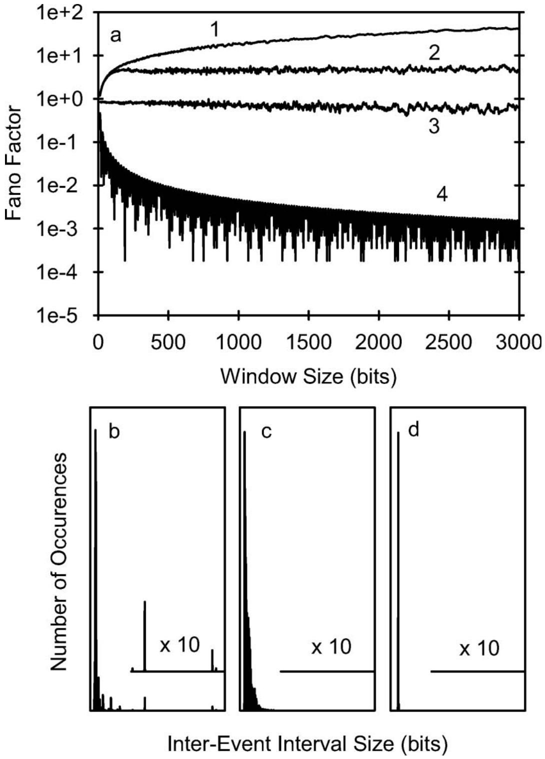


Fig. 2. (a) Fano factor calculations of simulated ~100,000-bit sequences of which 4–5% are 1s, of a fractal sequence (1), a fractal sequence in which the inter–event intervals are shuffled (2), a random sequence (3) and a periodic sequence (4). Each sequence is accompanied by an inter-event interval (IEI) histogram: (b) the IEI histogram of the simulated fractal sequence and of the shuffled IEI sequence. Both sequences have the same histogram, as the distribution of IEI is identical, despite the fact that their order of appearance has been altered. (c) The IEI histogram of the simulated random sequence. (d) The IEI histogram of the simulated periodic sequence.

size for artificially simulated ~100,000-bit sequences of a renewal process, a periodic process and a fractal process.

In a fractal process, the increase in variance occurs because clusters of long IEIs, which are characteristic of fractal processes, are more apt to be found as more and more data are collected. This increase can, on one hand, reflect a statistical self-similarity and thus, long-duration temporal correlations among the events, due to their order of appearance. On the other hand, it can result from the distribution of IEIs. The actual existence of long-duration temporal correlations within a given sequence is tested by randomly shuffling the data. This procedure creates a shuffled sequence whose mean IEI is identical to the mean IEI of the original sequence; however the IEIs are no longer correlated, because their order of appearance has been altered. The elimination of the increase of the Fano factor (as the window size increases) due to the shuffling procedure proves that long-duration temporal correlations existed in the original sequence. Therefore, the phenomenon is fractal.

We must point out that two major forms of fractal point processes can be synthesized: firstly, a "pure fractal point process" in which a Lévy distribution (Sato, 2004) is used to calculate the size of each IEI; and secondly, a "dependent fractal point process", in which a Lévy distribution is also used to calculate the size of each IEI together with a dependency factor that links the size of a certain IEI to the size of the preceding IEIs. In both fractal processes the Fano factor increases as a function of the window size. However, for a pure fractal process the increase of Fano factor as a function of the window size after shuffling will remain identical to the increase of the Fano factor as a function of the window size before shuffling. For the dependent fractal point process, the shuffling procedure will eliminate the increase of the Fano factor as a function of window size. In the Appendix, we explain the algorithm of synthesis of the fractal point processes of both classes of fractal processes and the effect of the data-shuffling procedure on the dynamics of the Fano factor.

THE USE OF ANAPHORIC CO-REFERENCES AS AN ENCODING KEY

In order to characterize a text as a specific point process (out of the aforementioned processes), one must first "binarily encode" the given text. Then, by applying the same statistical methods applied to electrical spike trains of neuronal networks to the text, one can characterize which type of point process the given text resembles. The encoding key, according to which the text should be encoded, must be inherent to the entity referred to as "text". That is, the encoding key should thread all the way through the given set of sentences to turn it into one body of information (i.e. text), on which the statistical methods can be applied.

The co-reference between individuals, events or theses within a given text is a prominent feature that enables us to establish a series of sentences as a text: namely, as one body of information, and thus determine the coherence of this set of sentences. The co-referents represent a certain motif that is repeated throughout the text. This repetition creates a sense of cohesion and correlation between the various parts of a given discourse. The notion of co-reference can be expressed either intertextually (endophoric reference) or extra-textually (exophoric reference). As we intend to apply quantitative methods to investigate a text, we will restrict ourselves to examining quantifiable motives, specifically the intertextual motives, and in particular the anaphora.

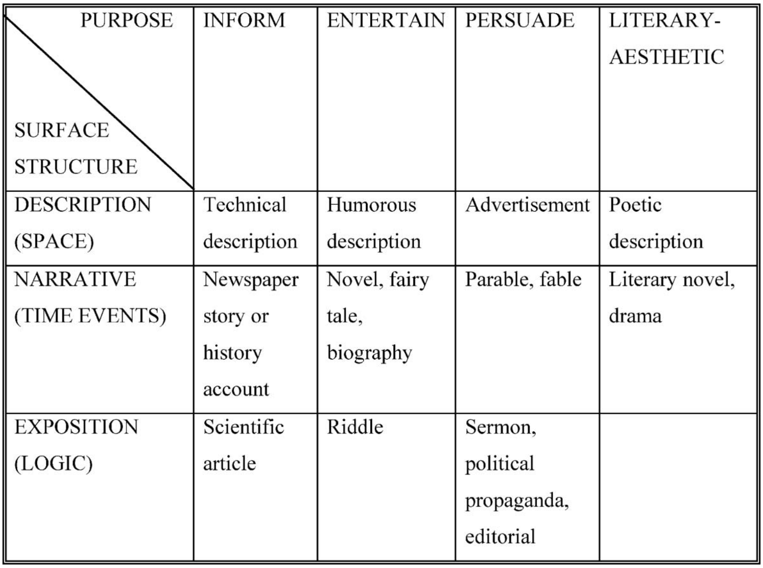
Anaphora is defined as a phenomenon consisting of the avoidance of redundancy, by the use of a semantically, lexically or phonologically attenuated expression in place of the initially complete lexical expression. By matching the anaphora with the initially used expression, the *antecedent*, the anaphora echoes the reference or the sense, which has already been established (Cornish, 1986).

In light of the above, a natural encoding key for the binary encoding of a given text is a list of anaphoric co-references.

THE CHOICE OF ANALYSED TEXT AND THE BINARY ENCODING PROCEDURE

In order to demonstrate the method to model a text as a specific kind of a point process (random, periodic or fractal), one must choose a long text, because the method we use is statistical. The text chosen must also resemble a time sequence since the method of analysis is customarily used to study the statistical properties of temporal behaviour of biological systems.

Based on these two requirements and on Brewer's well-accepted text classification scheme (Brewer, 1980), detailed in Figure 3, it was



PURPOSE SURFACE STRUCTURE	INFORM	ENTERTAIN	PERSUADE	LITERARY-AESTHETIC
DESCRIPTION (SPACE)	Technical description	Humorous description	Advertisement	Poetic description
NARRATIVE (TIME EVENTS)	Newspaper story or history account	Novel, fairy tale, biography	Parable, fable	Literary novel, drama
EXPOSITION (LOGIC)	Scientific article	Riddle	Sermon, political propaganda, editorial	

Fig. 3. Brewer's psychological classification of written discourse types.

reasonable to choose a long novel of the narrative genre. We have chosen Jonathan Swift's *Gulliver's Travels* (Swift, 1735) as our case study.

This novel is written in the form of a personal report. A personal report is typically characterized by a certain entity, the author of the report, which repeatedly appears throughout the text as a character rather than as just an omniscient narrator. The re-occurrence of such an entity is responsible for the cohesion of the sentences and consequently for the coherence of the entire text. In this specific text, the person who obeys these requirements is the imaginary Lemuel Gulliver.

Gulliver's Travels has been binarily encoded to create a sequence. This sequence is in the form of the electrical spike sequence (previously described), which is typically formed out of the continuous form of the neuronal membrane voltage.

We have encoded the ~100,000-word text by applying a carefully-designed key, according to which we could turn the series of phrases into a binary sequence. In Table 1 we present an arbitrarily-chosen paragraph of this text (which serves as a prelude to the first chapter in "The Voyage to Laputa", the third book in *Gulliver's Travels*), together with the key used to binarily encode it and the resulting binary sequence.

Table 1. A demonstrative example of the binary encoding procedure of text using the following encoding key: 1 = Lemuel Gulliver, I, me, my (NP), mine, 0 = any other word.

1	0	0	0	0	0	0	0	0	0	0	0
I	had	not	been	at	home	about	ten	days	when	captain	William

0	0	0	0	0	0	0	0	0
Robinson	a	Cornish	man	Commander	of	Hope-well	a	stout

0	0	0	0	0	0	0	1	1	0
ship	of	three	hundred	tons	came	to	my house	I	had

0	0	0	0	0	0	0	0	0	0
formerly	been	surgeon	of	another	ship	where	he	was	master

0	0	0	0	0	0	0	0	0	0	0
and	a	fourth	part	owner	in	a	voyage	to	the	Levant

0	0	0	0	1	0	0	0	0	0	0
He	had	always	treated	me	more	like	a	brother	than	an

0	0	0	0	0	1	0	1	0	0
inferior	officer	and	hearing	of	my arrival	made	me	a	visit

The encoding key that we have chosen is composed of four items: "I, me, my, mine". These four items are classified, by Halliday and Hasan (1976), as a single system of personal reference items, although they belong to different grammatical classes. In *Gulliver's Travels*, which is a personal report, direct speech does not appear. Hence, ambiguities as regards the identity of the antecedent of the key are avoided: namely, in this text the pronouns "I" and "me" always refer to the same character, Lemuel Gulliver. The possessive modifier "my", however, cannot be anaphorically related to these pronouns since it precedes objects other than Gulliver. Therefore, "my" can be considered a linking element and as such is bound to "I", "me" and "mine". In other words, the choice of items that constitute the encoding key is not arbitrary, as these items function as a single set of cohesive devices.

RESULTS OF THE ENCODING PROCEDURE OF LITERARY TEXTS AND DATA ANALYSIS

As explained in the previous paragraph, our analysis is devoted to written monologues. Figure 4 presents the Fano factor calculations as a function of the window size, for the original sequence and the shuffled IEIs' sequence of written monologues in various languages. *Gulliver's Travels*, which is presented in Figure 4(a), is one of them. Figure 4 (b–e) presents the results of the same analysis on other written monologues in English, French, German, and Hebrew, respectively. All chosen texts have approximately the same length, and thus produce sequences of similar lengths, $\sim$100,000 bits, except for the Hebrew text which is of $\sim$13,000 bits only (long [$\sim$100,000] Hebrew monologues are relatively few: those, which exist are rather new, therefore not open-access material yet). The keys used to encode each sequence are listed in Table 2. These keys are composed of phrases of an identical meaning (in each language) to the key used for *Gulliver's Travels*. The fraction of 1s in each sequence is approximately the same fraction (4–6%), except for the Hebrew text in which $\sim$12% are 1s.

We should point out, that (fortunately) in English, French and German, the existential "I" and "me" and the possessive "mine" act as one-word noun phrases (NPs). However, the possessive modifier "my" would always be accompanied by a noun, and often by additional modifiers. The number of additional modifiers is, in principle, infinite.

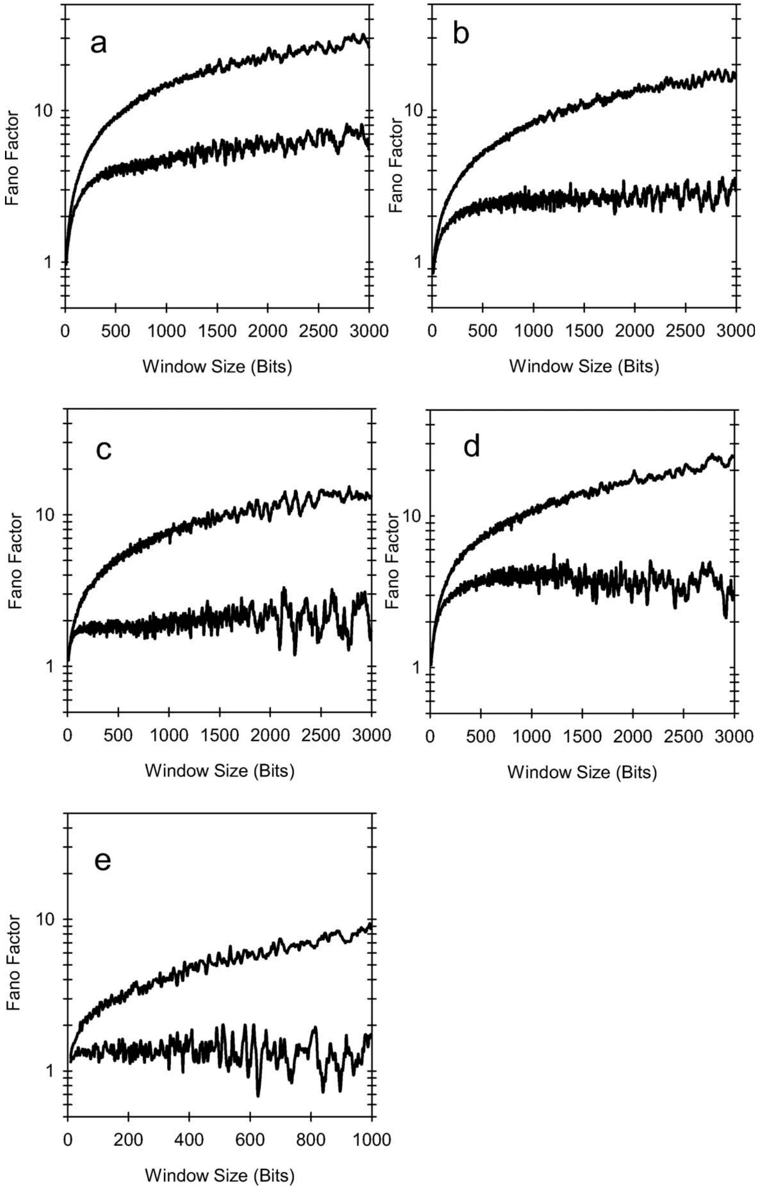


Fig. 4. Fano factor calculations of ∼100,000-bit sequences of written monologues in four different languages of which 46% are 1s. Fano factor calculations were performed on the original sequences as well as on the randomly shuffled IEI sequences. (a) English: *Gulliver's Travels*, (b) English: *Robinson Crusoe*, (c) French: *A Journey to the Centre of the Earth*, (d) German: *My Life and Death*, (e) Hebrew: *After-growth*.

Table 2. The literary texts used and the accompanying encoding keys.

Language	Text	Encoding key
English	*Gulliver's Travels* (Swift, 1735)	I, me, my (NP), mine
English	*Robinson Crusoe* (Defoe, 1719)	I, me, my (NP), mine
French	*A Journey to the Centre of the Earth* (Verne, 1864)	Je, me, moi, mon (NP), ma (NP), mes (NP), le mien, la mienne, les miens, les miennes
German	*My Life and Death* (May, 1910)	Ich, mich, mir, mein (NP), meinen (NP), meinem (NP), meine (NP), meiner (NP), meines (NP), mein, meine, meins
Hebrew	*After-growth* (Bialek, 1910)	Selected words ended by the letter "Yod"

Still, in the texts we have chosen, the most common possessive NPs are NPs that contain two additional modifiers. Therefore, any possessive modifier, referred to in the encoding keys, is regarded as a four-word NP: namely, the possessive modifier and the following three words are encoded as a single 1. We are, however, aware that there also exist possessive NPs with a different number of modifiers. Nevertheless, the analysed texts are long with respect to the fraction of phrases that conform to the encoding keys (as mentioned above). Therefore, we need (for the sake of analysis) only the approximate positions of the possessive NPs. In other words, there is a scarcely visible difference in the statistical results when using a possessive three-word, four-word or five-word NP.

In literary Hebrew, the encoding procedure is somewhat different, as one word can act as a phrase that contains both the possessive modifier and the accompanying noun. When a further modifier accompanies this one-word possessive NP, it begins with the definite article. In principle, we could have encoded selected words ended by the letter "Yod" and all possessive NPs composed of a word ended by the letter "Yod" and the following word preceded by the definite article, as a single 1. But, in this particular Hebrew text almost only one-word possessive NPs appear. Therefore, selected words ended by the letter "Yod" were encoded as 1s.

Each original sequence was constructed as explained in the previous paragraph. Each shuffled sequence was constructed out of the original sequence, by shuffling the order of the IEIs *only*. For reasons of clarity, the histogram of the IEIs of each original sequence and its accompanying shuffled sequence are not shown. However, if the IEI histograms of each original sequence and the associated shuffled sequence were presented, they would naturally be identical. We compared the Fano factor calculation results of each original literary sequence and its accompanying shuffled sequence (in Figure 4), with the Fano factor calculations of the simulated fractal sequence and the simulated shuffled fractal sequence (presented in Figure 2). We could, therefore, reliably deduce that regardless of the language, each original literary sequence reveals its self-similarity and thus long-duration correlations, as expected from a fractal process. This long-duration correlation is eliminated as the original sequence is shuffled.

At this time we may ask two questions: is the choice of phrases, of which the key is composed, and by which the text is encoded, responsible for the fractal structure of the text, or will any combination of phrases do? Is it possible that the fractal structure is the outcome of the syntactic roles of the phrases chosen for a specific key? To answer these questions, we encoded the same sample text, *Gulliver's Travels*, by using a different key: the "I, him, them, us, their (NP), his (NP), ours" key. The phrases in this alternative key play the same syntactic roles as the phrases in the "I, me, my (NP), mine" key. However, weaker rhetorical relations exist among the phrases in the "I, him, them, us, their (NP), his (NP), ours" key than among the phrases in the "I, me, my (NP), mine" key. The results of the Fano factor calculations of the sequence formed by using the "I, him, them, us, their (NP), his (NP), ours" key are presented in Figure 5. The fraction of 1s in the sequence created by using the "I, him, them, us, their (NP), his (NP), ours" key is similar (4.49%) to the fraction of 1s in the sequence created by using the "I, me, my (NP), mine" encoding key (5.42%). The Fano factor calculation of the sequence formed using the "I, him, them, us, their (NP), his (NP), ours" key resembles a renewal point process more than a fractal process. Hence, one can carefully infer that the fractal structure of the given text does not stem out of the syntactic roles of the phrases that create the encoding key, as previously hypothesized.

This test, tentatively, provides evidence that in order to characterize a text as a fractal system, certain rhetorical relations among the phrases that compose the encoding key should exist.

One can justifiably claim that eliminating the effect of syntactic roles is not a proof that it is the rhetorical relations among the phrases composing the encoding key that affect the fractal structure. Alternatively, another explanation could account for the decrease in the Fano factor as a function of window size, using a text encoded by the "I, him, them, us, their (NP), his (NP), ours" key: the "I, him, them, us, their (NP), his (NP), ours" key is an encoding key which is a superposition of several incomplete keys. Thus, this

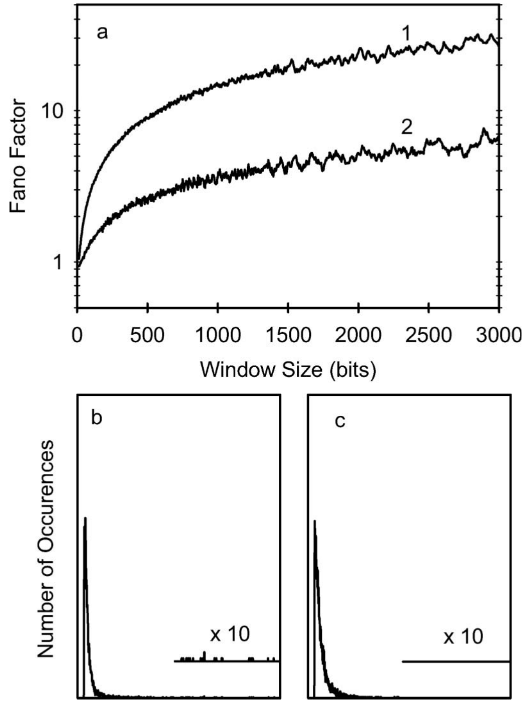


Fig. 5. (a) Fano factor calculations of two sequences formed out of *Gulliver's Travels* by using two different keys with approximately the same number of events (~ 5%): (1) "I, me, my (NP), mine", (2) "I, him, them, us, their (NP), his (NP), ours". (b) The long-tailed inter-event interval histogram of the original key "I, me, my (NP), mine". Note that events occur even on long time scales. (c) The inter-event interval histogram of the encoding key "I, him, them, us, their (NP), his (NP), ours". Note that no events occur on long time scales.

encoding procedure creates a sequence, which is a superposition of several partial sequences. The fractal structure is not fully recovered, since a certain fraction of 1s has been artificially excluded from the calculation and a certain fraction of 1s artificially included. This can also explain the decrease of the Fano factor as a function of window size.

Whether one accepts "the lack of rhetorical relations" argument or prefers "the incomplete encoding keys" argument, they both stand for the fact that in order to reveal the fractal structure of a given text, a complete set of phrases should be used as a key. This complete set should consist of all the phrases among which rhetorical relations exist. If such a complete set is not used, but rather a partial set or a superposition of several sets, then the fractal structure of the text cannot be fully revealed.

We have performed identical tests on the other four texts that appear in Figure 4, and obtained the same results irrespective of the language in which the text is composed.

An alternative approach to examine the effect of syntactic roles of the phrases of which the key is composed, is to check what affects the order of the IEIs series: is it the order of IEIs within a sentence, namely the internal structure of the sentence, or is it the IEIs across sentences? This examination is equivalent to asking whether the fractal structure depends on a collection of sentences that form one body of information, namely a text, or if any random collection will do.

In Figure 6 we present an answer to this question. Again, we show the Fano factor calculations of the original sequence of *Gulliver's Travels* using the "I, me, my (NP), mine" key, and then the Fano factor calculations on sequences in which:

a. The phrases were randomly shuffled all over the text (2).
b. The IEIs were randomly shuffled (3).
c. The sentences were shuffled without changing the order of phrases in each sentence (4).

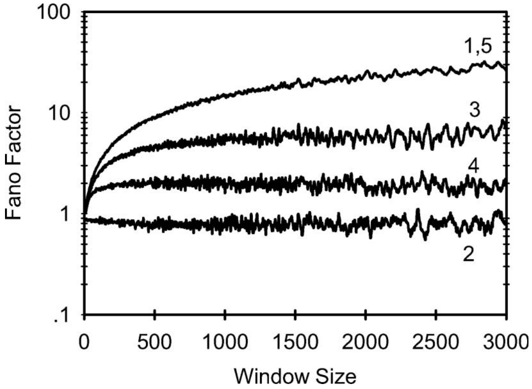


Fig. 6. Fano factor calculations of five sequences constructed out of the binary sequence of *Gulliver's Travels* using the "I, me, my (NP), mine" key: (1) the original sequence, (2) a sequence in which the phrases were randomly shuffled, (3) a sequence in which the inter-event intervals were randomly shuffled, (4) a sequence in which the sentences were randomly shuffled but the order of phrases within a sentence is the original order, (5) a sequence in which the order of sentences is the original order of sentences, but the order of phrases within those sentences is randomly shuffled.

d. The sentences were kept in place while shuffling the order of phrases within each sentence (5).

"Shuffling the order of phrases all over the text" is shuffling the order of events, which naturally eliminates the fractal structure. "Shuffling the IEIs", which has already been presented in Figure 4, proves that the original sequence is fractal in nature. "Shuffling the sentences without changing the order of the phrases" within the sentences considerably reduces the Fano factor calculation. However, the Fano factor calculations following "shuffling the phrases within each sentence, without changing the order of sentences" greatly resemble the Fano factor calculation of the original sequence.

The shuffling of phrases within a sentence is a simulation of changing the syntactic roles of the phrases within that sentence. This kind of shuffling slightly changes the meaning of the sentence but one can still understand its essence. The fact that this procedure has hardly any discernible effect on the Fano factor calculations proves that the syntactic roles of phrases have only a minor effect on the calculations. (This has also been previously presented). It can, however, be rationalized on the basis of the fact that the analysis we use is a low-resolution statistical technique and the binary sequences we use are rather large. In a 100,000-bit vector, the change of the location of a word within a sentence is rather minor. Therefore, this intra-sentence mixing will have a minor effect on our results.

Both explanations support, however, the conclusion that the order of sentences within a given text is what accounts for the significant Fano factor calculation results of the original sequence.

SUMMARY AND CONCLUSIONS

We have binarily encoded written monologues in a similar manner to the encoding of a continuous electrical signal of neuronal networks. The encoding key has been carefully designed to include phrases that are co-referents. Only such encoding keys produce binary series, which have a fractal structure: namely, they are self-similar and thus long-duration correlated. This fractal structure can be revealed by using a statistical tool named "Fano factor".

By outlining the analogy between the structure of long-duration neuronal electrical activity and the structure of text, one can account for the hierarchical organization of natural texts: the hierarchical structure of text can be regarded as a reflection of the structure of the electrical activity of neuronal networks, in order to facilitate data processing within the brain. This phenomenon can be considered as a manifestation of the ‘‘form follows function’’ concept.

This conclusion recalls the ideas of textual hierarchical organization developed by W. C. Mann and S. A. Thompson in the framework of rhetorical structure theory – RST (Mann & Thompson, 1988). They showed that hierarchical analysis of text is independent of the morphology of syntactic signals, but depends on semantic judgments between groups of clauses. These ideas had been functionally formulated in terms of the purposes of the writer and writer's assumptions about the reader (Mann & Thompson, 1988, p. 270):

> In recognizing text structure, the reader adds structure to a linear string. ... If we see part of the function of communication as building memories, we can see nuclearity as suggesting organizational details of those memories. If the text structure, even in part, represents the access patterns that are facilitated in memory, then nuclearity can be seen as a way to signal that the memory of a satellite can be accessed through the nucleus. ...

Based on these conclusions, it would now be interesting to use the Fano factor calculation not only for analytical purposes, but also as a constructive tool for text generation.

ACKNOWLEDGMENTS

We thank Vladislav Volman for helpful discussions at the early stages of this research, Yonatan Becker for his valuable remarks while reading this manuscript, and Alan J. Atlas for proofing this manuscript.

REFERENCES

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-organized criticality: An explanation of 1/f noise. *Physical Review Letters*, *59*, 381–384.

Bak, P., Tang, C., & Wiesenfeld, K. (1988). Self-organized criticality. *Physical Review*, *A38*, 364–374.

Bialek, H. N. (1910). *Safiach*, 1953 edition. Tel-Aviv: Dvir.

Brewer, W. F. (1980). Literary theory, rhetoric and stylistics, In R. J. Shapiro, B. C. Bruce & W. F. Brewer (Eds), *Theoretical Issues in Reading Comprehension* (pp. 221–239). Hillside, New Jersey: Erlbaum.

Cornish, F. (1986). *Anaphoric Relations in English and French*. Dover, New Hampshire: Croom Helm.

Cox, D. R., & Lewis, P. A. W. (1996). *The Statistical Analysis of Series of Events*. London: Chapman and Hall.

Defoe, D. (1719). *Robinson Crusoe*. London: William Taylor.

Fano, U. (1947). Ionization yield of radiations. II. The fluctuations of the number of ions. *Physical Review*, *72*, 26–29.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Janicki, A., & Weron, R. (1994). *Simulation and Chaotic Behaviour of α-Stable Stochastic Processes*. New York: Marcel Dekker.

Lowen, S. B., & Teich, M. C. (1996). The periodogram and Allan variance reveal fractal exponents greater than unity in auditory-nerve spike trains. *Journal of the Acoustics Society of America*, *99*, 3585–3591.

Lowen, S. B., Cash, S. S., Poo, M.-M., & Teich, M. C. (1997). Quantal neurotransmitter secretion rate exhibits fractal behaviour. *Journal of Neuroscience*, *17*, 5666–5677.

Lewis, C. D., Gebber, G. L., Larsen, P. D., & Barman, S. M. (2001). Long-term correlations in the spike trains of medullary sympathetic neurons. *Journal of Neurophysiology*, *85*, 1614–1622.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281.

May, K. F. (1910). *Mein Leben und Sterben*. Freiburg: Verlag von Friedrich Ernst Fehsenfeld.

Powers, N. L., & Salvi, R. J. (1992). Comparison of discharge rate fluctuations in the auditory nerve of chickens and chinchillas. In *Abstracts of the XIV Midwinter Research Meeting*. (p. 101). Association for Research Otolaryngology, Des Moines, Iowa.

Powers, N. L., Salvi, R. J., & Saunders, S. S. (1992). Discharge rate fluctuations in the auditory nerve of the chinchilla. In *Abstracts of the XIV Midwinter Research Meeting* (p. 101). Association for Research Otolaryngology, Des Moines, Iowa.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1992). *Numerical Recipes in C*. New York: Cambridge University Press.

Saiki, T., Kitagawa, Y., & Hayashi, A. (1999). Fluctuations of character centroid intervals in laterally written Japanese sentences. *IEICE Transaction Fundamentals*, *E82-A* (3), 520–526.

Sato, K.-I. (2004). *Levy Processes and Infinitely Divisible Distributions* (Cambridge Studies in Advanced Mathematics). Cambridge: Cambridge Press.

Schmeisser, E. T., McDonough, F. J. M., Bond, M., Hislop, P. D., & Epstein, A. D. (2001). Fractal analysis of eye movement during reading. *Optometry and Vision Science*, *78*(11), 805–814.

Swift, J. (1735). *Gulliver's Travels*. London: George Routledge and Sons (1906 edition edited by Henry Morely).

Teich, M. C., Heneghan, C., Lowen, S. B., Izaki, I., & Kaplan, E. (1997). Fractal character in the neural spike sequence in the visual system of the cat. *Journal of the Optical Society of America*, *A14*, 529–546.

Teich, M. C., Turcott, R. G., & Siegel, R. M. (1996). Temporal correlation in cat striate-cortex neural spike trains. *IEEE Engineering in Medicine and Biology*, *15*(5), 79–87.

Turcott, R. G., & Teich, M. C. (1993). Long-duration correlation and attractor topology of the heartbeat rate differ for healthy patients and those with heart failure. *Proceedings of SPIE 2036* (Chaos in Biology and Medicine, pp. 22–39).

Verne, J. (1864). *Voyage au Centre de la Terre*. Bibliothèque d'Éducation et de Récréation J. Hetzel et Cie.

Weron, R. (1996). On the Chambers-Mallows-Stuck method for simulating skewed stable random variables. *Statistics and Probability Letters*, *28*, 165–171.

APPENDIX

The central limit theorem states that a sum of N independent and identically distributed random variables X_n, with finite first and second moments, obeys a Gaussian distribution in the limit $N \to \infty$. That is, if $P_n \equiv \sum_{n=1}^{N} X_i$ is the partial sum of the above random variables, then the central limit theorem holds and is distributed as a Gaussian in the limit when $N \to \infty$.

Many, but not all, statistical distributions belong to the domain of attraction of the Gaussian. There is a whole class of distributions that do not fulfil the hypothesis of a finite second moment, such as:

$$P(x) \sim \frac{1}{|x|^{1+\alpha}} \quad \text{where} \quad |x| \to \infty \quad \text{and} \quad 0 < \alpha < 2$$

This kind of inverse power-law tail precludes the convergence to the Gaussian distribution but not the existence of a limiting distribution. In general, under some criteria, the distributions that obey the above power-law tail are known as Lévy α-stable distributions. One of the most prominent features of the Lévy distributions is that they exhibit fractal behaviour of stochastic processes, over time. Such processes are known as Lévy processes.

In order to synthesize a fractal-like distributed vector that is compatible with our binary encoded text vector, we used a function that evaluates a random number generated by a Lévy α-stable distribution (Janicki & Weron, 1994; Weron, 1996).

The algorithm for constructing a standard stable random variable $X \sim S_\alpha(1, \beta, 0)$ where $0 < \alpha \leq 2$ and $-1 \leq \beta \leq 1$ is the following:

- Generate a random variable Θ uniformly distributed on $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and an independent exponential random variable W with mean 1.

- In the nonsymmetrical case, where $\beta \neq 0$
  - ▪ For $\alpha \neq 1$
    Define:

$$\Phi_{\alpha,\beta} = \arctan\,[\beta \cdot \tan(\pi\alpha/2)]/\alpha$$
$$S_{\alpha,\beta} = [1 + \beta^2 \cdot \tan^2(\pi\alpha/2)]^{1/2\alpha}$$

Then:

$$X = S_{\alpha,\beta} \bullet \left[\frac{\sin\left[\alpha \cdot \left(\Theta + \Phi_{\alpha,\beta}\right)\right]}{\left[\cos(\Theta)\right]^{1/\alpha}}\right] \bullet \left[\frac{\cos\left[(1-\alpha)\cdot\Theta + \alpha\Phi_{\alpha,\beta}\right]}{W}\right]^{(1-\alpha)/\alpha}$$

  - ▪ For $\alpha = 1$

$$X = \frac{2}{\pi}\left[\left(\frac{\pi}{2} + \beta\Theta\right)\tan\Theta - \beta \cdot \ln\left(\frac{\frac{\pi}{2}\cdot W \cdot \cos\Theta}{\frac{\pi}{2} + \beta\Theta}\right)\right]$$

- In the symmetrical case, where $\beta = 0$ the above expressions are reduced to:
  - ▪ For $\alpha \neq 1$

$$X = \left[\frac{\sin(\alpha\Theta)}{\left[\cos(\Theta)\right]^{1/\alpha}}\right] \bullet \left[\frac{\cos[(1-\alpha)\cdot\Theta]}{W}\right]^{(1-\alpha)/\alpha}$$

  - ▪ For $\alpha = 1$ $X = \tan\Theta$

The random generator function can be given an independent uniform (0,1) random number U using the ran2 function of *Numerical Recipes* (Press et al., 1992). This function can get up to $2{\cdot}10^{18}$ random numbers without correlations. It is easy to obtain Θ and W from independent uniform (0, 1) random variables U_1 and U_2 by setting $\Theta = \pi\ (U_1 - ½)$ and $W = -\log\ (U_2)$.

To synthesize the fractal distribution vector with autocorrelation between the inter-event intervals, we used the above function to calculate the uncorrelated random part of the IEI. However, the correlated part of the IEI was calculated from the previous N IEIs. We mathematically represent the expression for the Kth IEI as:

$$D_k = R_k(\alpha) \quad \text{for} \;\; 0 \leq k < N$$

$$\begin{aligned} D_k = {} & P_0 D_{k-N} + P_1 D_{k-N+1} + P_2 D_{k-N+2} \\ & + \cdots + P_{N-1} D_{k-1} + P_N R_k(\alpha) \quad \text{for} \;\; k \geq N. \end{aligned}$$

where $R_k(\alpha)$ is a Lévy random number and P_i is a probability for each component in the sum. Obviously, $\Sigma P_i = 1$.

From this description it is obvious that in an IEI distribution, which obeys a pure Lévy distribution, the IEIs are totally independent, as just drawing lots produced them. Hence, the shuffling procedure has no effect on any statistical end-result, as the shuffling procedure is equivalent to drawing lots again.

The shuffling procedure, however, has a major effect on statistical quantities when it comes to a dependent Lévy distribution as the value of each IEI depends to some extent on the values of the previous IEIs. In the case of a dependent Lévy distribution, the shuffling procedure eliminates the mutual dependence among IEIs. If the weight of dependence is considerable, then the shuffling procedure results in a random Gaussian distribution rather than a Lévy distribution. This evidently changes the outcome of certain statistical quantities such as the Fano factor.