Physica A 387 (2008) 6411–6420

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Equilibrium and dynamic methods when comparing an English text and its Esperanto translation

M. Ausloos

GRAPES, U. Liege, B5 Sart-Tilman, B-4000 Liege, Belgium

ARTICLE INFO

Article history:
Received 26 February 2008
Received in revised form 27 May 2008
Available online 23 July 2008

Keywords:
Text
Language
Translation
Zipf
Grassberger–Procaccia
Time series

ABSTRACT

A comparison of two English texts written by Lewis Carroll, one (Alice in Wonderland), also translated into Esperanto, the other (Through the Looking Glass) are discussed in order to observe whether natural and artificial languages significantly differ from each other. One dimensional time series like signals are constructed using only word frequencies (FTS) or word lengths (LTS). The data is studied through (i) a Zipf method for sorting out correlations in the FTS and (ii) a Grassberger–Procaccia (GP) technique based method for finding correlations in LTS. The methods correspond to an equilibrium and a dynamic approach respectively to human texts features. There are quantitative statistical differences between the original English text and its Esperanto translation, but the qualitative differences are very minutes. However different power laws are observed with characteristic exponents for the ranking properties, and the *phase space attractor dimensionality*. The Zipf exponent can take values much less than unity ($\sim$0.50 or 0.30) depending on how a sentence is defined. This variety in exponents can be conjectured to be an intrinsic measure of the book style or purpose, rather than the language or author *vocabulary richness*, since a similar exponent is obtained whatever the text. Moreover the attractor dimension r is a simple function of the so called phase space dimension n, i.e., $r = n^{\lambda}$, with $\lambda = 0.79$. Such an exponent could also be conjectured to be a measure of the author *style versatility*, — here well preserved in the translation.

1. Introduction

Human written languages are systems usually composed of a large number of internal components (the words, punctuation signs, and blanks in printed texts) which obey rules (grammar) [1,2]. Relevant questions pertain to the life time, concentration, distribution,.. complexity of these and their relations between each others. Thus human language is a new emerging field for the application of methods from the physical sciences in order to achieve a deeper understanding of linguistic complexity [3–9].

One should distinguish two main frameworks. On one hand, language developments seem to be understandable through competitions, like in Ising models, and in self-organized systems. Their diffusion seems similar to percolation and nucleation-growth problems taking into account the existence of different time scales, for inter- and intra-effects. The other frame is somewhat older and originates from more classical linguistics studies; it pertains to the content and meanings [1,2]. This latter case is of interest here and the main subject of the report, within a statistical physics framework.

Concerning the internal structure of a text, supposedly characterized by the language in which it is written, it is well known that a text can be mapped into a signal, of course first through the alphabet characters. However it can be also reduced to less abundant symbols through some threshold, like a time series, which can be a list of $+1$ and -1, or sometimes 0. Thereafter one could apply at this stage many techniques of signal analysis.

E-mail address: marcel.ausloos@ulg.ac.be.

0378-4371/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2008.07.016

In fact, laws of text content and structures have been searched for a long time by e.g. Zipf and others [10–14] through the least effort (so-called ranking) method. It has been somewhat a surprise that the number of words $w(h)$ which occurs h times in a text is such that $w(h) \sim 1/h^{\gamma}$, where $\gamma \sim 2$, while the rank R of the words according to their frequency f behaves like another power law $f \sim R^{-\zeta}$ where the exponent ζ is quasi always close to 1.0 [15,16], due to constrained correlations [17–19]. One can also show that $w(h) \sim 1/h^{1+\nu}$ [20]; whence $\gamma = 1+\nu$, with $\nu = 1/\zeta$ [21,22].

Another distribution can be studied, i.e. the distribution of word lengths in a text, e.g. Refs. [23,24]. Whence two features can be looked for (i) word frequencies (FTS) or (ii) word lengths (LTS). The first one leads to characterizing the spanned phase space through a measure, — it is a static-like, equilibrium approach, obtained *after* the text is finalized, while the second rather contains a time *evolution* aspect: it takes more time to write (or pronounce or read) a long word than a small one. The Grassberger–Procaccia (GP) technique can be implemented for finding "time correlations" in the text through the analysis of LTS, as a signal spanning some attractor in a space on an *a priori* unknown dimension.

Obviously there are many ways to map a text onto a time series, but in the present study the above two series are only considered, due to their physical meaning which can be thought to be implied in the mapping.

"Linguistic time series" have often studied at a letter or word level [25–28] or as in Montemurro and Pury [27,28] at a frequency mapping, similar though not identical to the one described below. Others have considered Zipf law(s) at the sentence level [29,30].

Esperanto is an artificially and somewhat recently constructed language [31], which was intended to be an easy-to-learn lingua franca. Previous statistical analyses seem to indicate that Esperanto's statistical proportions are similar to those of other languages [32]. It was found that Esperanto's statistical proportions resemble mostly those of German and Spanish, and somewhat surprisingly least those of French and Italian. By the way, English seems to be an intermediary case [14]. Yet there are quantitative differences: English contains ∼1 M words [33], esperanto 150 k words [34]. Other artificial (spoken) languages exist, like that of the Magma [35] and Urban Trad [36] music groups, the latter specifically designed for song competition. Moreover, as in e.g. rap music lyrics or French *verlan*, the thesaurus is rather of limited size in all these cases.

To my knowledge few comparisons exist on texts translated from one to another language [37–39], in particular into artificial languages. An original consideration is presented here below, i.e. the analysis and results about a translation between one of the most commonly used language, i.e. English, and a relatively recently *created* language, i.e. esperanto.

The text to be used was chosen for its wide diffusion, freely available from the web [40] and as a representative one of a famous scientist, Lewis Caroll, i.e. *Alice in wonderland* (AWL) [41]. Knowing the special (mathematical) quality of this author's mind, and expecting some, possibly special way of writing, another text has been chosen for comparison, i.e. *Through the Looking Glass* (TLG) [42]; — alas to my knowledge only available in English on the web [43]. Yet this set will allow one to discuss whether any observed difference between the Esperanto and English versions are due to the translation or on the contrary to the specificity of this author's creativity. It might be also expected that one could observe whether some style or vocabulary change has been made between two texts having appeared at different times: 1865 and 1871, or not. Previous work on the English AWL version should be mentioned [14], where a relevant ingredient to be taken into account in discussing most written texts is emphasized, i.e. a mixing of oral and descriptive accounts.

In Section 2, a few elementary facts and basic statistics on these texts are presented; the methodology is briefly exposed, i.e. as one recalls (i) two simple ways to map texts into *signals*, i.e., the frequency time series (FTS) and the word length time series (LTS), (ii) the Zipf ranking technique, (iii) the Grassberger–Procaccia (GP) method [44,45] used for finding correlations. Similar techniques for comparing English and Greek texts, but not from a translation point of view can be found in Ref. [23]; however the published work contains a few annoying (misprints or) defects whence the present reformulation of the techniques when applied to text problems. In Section 3, the results are presented: (i) a Zipf analysis on the frequency time series (FTS), (ii) a GP analysis for the word length time series (LTS). The results are discussed in Section 4.

2. Data and methodology

For these considerations two texts here above mentioned and one translation have been selected and downloaded from a freely available site [40], resulting obviously into three files. The chapter heads have been removed. All analyses are carried out over this reduced file for each text. Basic statistics, like the number of words, the longest sentence, ... are given in Table 1 for each text, and chapters. A few facts attract some attention

(1) the number of dots is *much smaller* in AWL_{eng} than in AWL_{esp} and also in TLG_{eng}
(2) automatically the longest sentence occurs in AWL_{eng} with many more characters
(3) the longest sentence in AWL_{esp} occurs between commas
(4) the number of semi-colons is very small in TLG_{eng}
(5) the longest sentence ever occurs in TLG_{eng} between semi-colons
(6) there are *very few* exclamation marks in AWL_{esp}
(7) but a long sentence is then found between these in such a work
(8) more importantly the number of sentences is much smaller in AWL_{eng} than in AWL_{esp}.

Let us now search for correlations in the texts through both ways of constructing a time series from such documents of e.g. M words:

Table 1
Basic statistical data for the three texts of interest; in each case the longest sentence is measured in terms of the number of characters (not in terms of words)

	AWL_{eng}	AWL_{esp}	TLG_{eng}
Number of words	27 342	25 592	30 601
Number of different words	2 958	5 368	3 205
Number of characters	144 927	154 445	16 4147
Number of punctuation marks	4 481	4 752	4 828
Number of "sentences"	1 633	2 016	2 059
Words in chap. 1	2 194	1 858	
Different words in chap. 1	652	853	
Words in chap. 2	2 188	1 915	
Different words in chap. 2	665	829	
Number of dots	979	1 545	1 315
Longest "sentence"	1 669	825	864
Number of commas	2 419	2 324	2 441
Longest "sentence"	373	1 170	368
Number of semi-colons	195	207	72
Longest sentence	6 624	6 043	12 501
Number of colons	234	205	256
Longest sentence	4 586	5 576	3 145
Number of question marks	203	205	254
Longest sentence	6 323	5 581	5 212
Number of exclamation marks	451	266	490
Longest sentence	4 388	6 249	4 016

(1) Count the frequency f of appearance of each word in the document. Rewrite the text such that at each "appearance" of a word, the word is replaced by its frequency such that one obtains a time series $f(t)$. Such a time series is called a "frequency time series" (FTS).
(2) Count the number l of letters of each word located in the text successively at the *time* $t = 1$, for the first word, at time $t = 2$, for the second, etc. Construct a time series $l(t)$. Henceforth, such a time series is called a length time series (LTS).

These two sorts of time series are thereby analyzed along one of the two mentioned techniques, one being more pertinent than the other as outlined here above. Let us discuss them briefly.

2.1. Zipf method

A large set of references on Zipf's law(s) in natural languages can be found in Ref. [46]; see also the journal *Glottometrics* vol. 2–5. The idea has been already applied to many various complex signals or "texts", — signals, translated through a number k of characters characterizing an alphabet, like, among many others, for time intervals between earthquakes [47], DNA sequences [48] or for financial data [49].

Zipf calculated the number N of occurrences of each word in a given text. By sorting out the words frequency f according to their rank one expects

$$f \sim R^{-\zeta} \tag{1}$$

with an exponent ζ close to unity, as recalled in the Introduction. This strong quantitative statement is attested over a vast repertoire of human languages [15]. Yet it is of empirical evidence that Zipf' s law in this (FTS) form can at most account for the statistical behaviour of words frequencies in a zone spanning the middle — low to low range of the rank variable. Even in the case of short texts Zipf s law renders an acceptable ζ in the small window between $s \simeq 10$ and 1000, which does not represent a significant fraction of any literary vocabulary.

One difficulty stems in the lower and upper ranks of such plots because of the abundance and rarity of words [50]. Mandelbrot [51–53] using arguments based on *fractal* ideas, applied to the structure of lexical trees, improved the original form of the law, writing, in terms of two parameters A and C that need to be adjusted to the data,

$$f(R) = \frac{A}{(1 + CR)^{\zeta^*}}. \tag{2}$$

It has been shown that this Zipf–Mandelbrot (Z–M) law is also obeyed by so many random processes [54,55] that one has been sometimes ruling out any interestingly special character for *linguistic* studies. Nevertheless, it has been argued that it is possible to discriminate between human writings [56] and stochastic versions of texts precisely by looking at statistical properties of words that fall where Eq. (1) does not hold [23]. Whence some question cannot be avoided on such law validity/deviation in artificial languages and on effects resulting from automatic or machine translations [39].

Note that the length of sentences is also going to be examined from the point of view of the numbers of characters between (six sorts of) punctuation marks, see Table 1. One also investigates whether the first and second chapter of AWL_{eng} and AWL_{esp} are "Zipfly" similar, thereby allowing for some consistency test.

Table 2
Top ten most frequent words in AWL$_{eng}$ and AWL$_{esp}$ with their frequency, indicating that a translator can or sometimes must modify the style and vocabulary, e.g. exchanging pronouns for nouns or conversely

AWL$_{eng}$	f	AWL$_{esp}$	f
the	1527	la (=the)	2070
and	802	kaj (=and)	628
to	725	ŝ i (=she)	508
a	615	ne (=no/not)	426
I	545	mi (=I)	403
it	527	Alicio (=Alice)	347
she	509	diris (=said)	332
of	500	al (=to)	313
said	456	vi (=you)	302
Alice	395	ke (=that)	292

2.2. Grassberger–Proccacia method

In order to get an insight into the dynamics of a system solely from the knowledge of the time series, a method derived by Grassberger and Proccacia [44,45] has been proven to be particularly useful. This method has been applied to analyze the dynamics of neural network activity electric activity of semiconducting circuits [57,58], climate [59], etc.

We aim to finding some answer to questions like

(1) Can the salient features of the system be viewed as the manifestation of a deterministic dynamics, or do they contain an irreducible stochastic element?
(2) Is it possible to identify an attractor in the system phase space from a given time series [60]?
(3) If the attractor exists, what is its dimensionality r [61]?
(4) What is the (minimal) dimensionality n of the phase space within which the above attractor is embedded [62]?

This is done as follows: Let the LTS time series having M data points, i.e. y_i $(i = 1, \dots, M)$. Consider the data as illustrating some dynamical process in a (phase) space with dimension n. Construct a set of V vectors v_k $(k = 1, \dots, V)$ containing $n-1$ points as follows:

$$v_k = (y_k, y_{k+\tau}, y_{k+2\tau}, \dots, y_{k+(n-1)\tau}) \tag{3}$$

where τ is an integer, called the *delay time*. Notice: $V + (n-1) = M$. In other words, one considers $k + n\tau$ as a sum modulo M. Next one estimates the correlation integral from the *distance* $|v_i - v_j|$ between all the vectors such that $1 \le i, j \le V$. The correlation integral $C_n(l)$ is obtained from

$$C_n(l) = \frac{\#\text{ of pairs } (i,j) \text{ such that } |v_i - v_j| < l}{N^2}. \tag{4}$$

In other words,

$$C_n(l) = \frac{\#\text{ of pairs } (i,j>i) \text{ such that } |v_i - v_j| < l}{N(N-1)/2}. \tag{5}$$

GP have shown that for small l, one has

$$C_n(l) \simeq Bl^r \tag{6}$$

where B is some constant and r is the so called *attractor (correlation) dimension*, measuring the number of dynamic variables or number of degrees of freedom. In order to obtain r for the different n values, a log–log plot is in order. The choice of τ is debatable [63]. Here $\tau = 500$ as in other related studies [23], for $n = 1$ to 15.

Practically, it was noticed that the correlation integral calculated for $|v_i < v_j|$ distances takes a finite number of values; therefore each distance l was "measured" up to three decimal digits. Therefore two distances differing by less than 0.001 are not differentiated. Even though the robustness of this "numerical approximation" has not been tested, it is likely not a drastic one.

A fit of the beginning of the $C_n(l)$ evolution through the best mean square technique on a log–log plot leads to a value of the relevant slopes, thus r defined by Eq. (6).

3. Results

3.1. Zipf plots: FTS analysis

The top ten most frequent words in AWL$_{eng}$ and AWL$_{esp}$ are given with their frequency in Table 2. It seems of interest to point out differences in style appearing from such a table. Notice that a translation does not conserve the number of words in a text, nor their importance in frequency. Of course the ranking might be intrinsically different, but also the translator

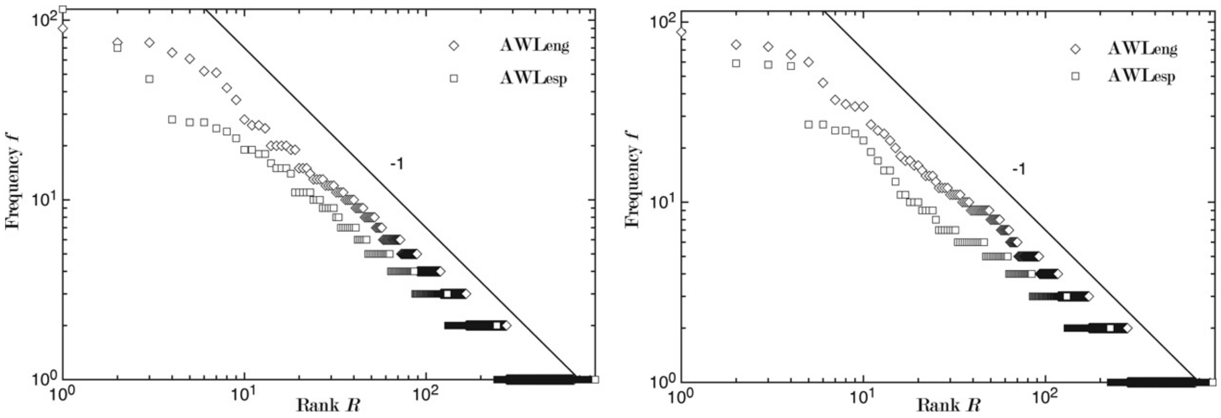

Fig. 1. Consistency of natural and artificial languages through written texts: Zipf (log–log) plot of the frequency of words in (a) chapter 1, (b) chapter 2, for two texts of interest, i.e. AWL_{eng} and AWL_{esp}. The "usual" 1 exponent is indicated.

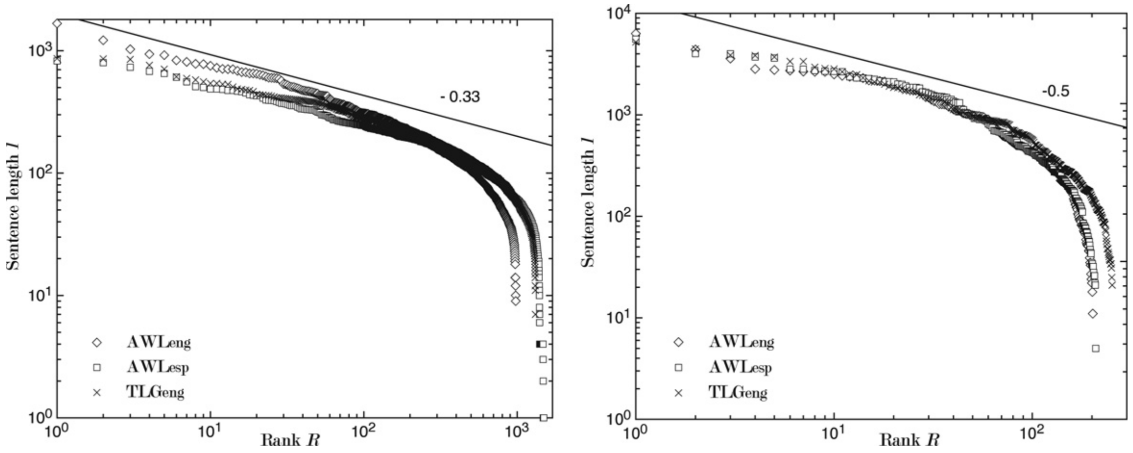

Fig. 2. Zipf (log–log) plot of the FTS of sentence lengths, as separated by (a) dots, (b) question marks, in the three texts of interest AWL_{eng}, AWL_{esp} and TLG_{eng}. The best exponent of the corresponding Zipf law is indicated as a guide to the eye.

can (or must) modify some sentences according to the language and grammar. The translator might use nouns rather than pronouns if these are confusing. An interesting illustration is noticed in Table 2: the same words *'the'= 'la'* and *'and'= 'kaj'* are both times the most frequent, but for example *'Alice'* occurs more frequently in the English text than in the Esperanto text, same for *'I'= 'mi'*, even though *'she'= 'ŝi'* occurs an equal number of times.

The (word frequency, rank) relation roughly follows a linear relationship, $\zeta \simeq 1$ both for the English and Esperanto texts as well as for chapter 1 and 2 of both AWL texts; the latter plots are shown in Fig. 1 for AWL. All the other rather unspectacular results are not shown, as requested by a referee; all plots of the FTS analysis for the three texts can be found in Ref. [64]. Some change in curvature is found for all texts at low R where a so called discontinuity exists; it is explained by [14,65] as due to a transition between colloquial ("common") small and "distinctive" words. Some break, or change in slope, is also found for $R \sim 100$, — see discussion also in Refs. [14,65]. Interestingly let it be pointed out that the Rank = 1 word has a much higher frequency in the Esperanto text than in the English texts. Moreover, the variety of distinct words is larger in Esperanto as well. In between, the number of Esperanto words is less frequent in general at a given rank indicating a greater simplicity in the Esperanto "ordinary" vocabulary.

A question not often investigated is the relation between characters, words, sentences, paragraphs, etc. Here one has investigated the length of sentences, defined through the usual separators, in the three texts, in a Zipf plot way. The number of punctuation marks is relatively equivalent (see Table 1) in all texts, but the number of dots and of commas are much larger than the other punctuation marks. Therefore one might expect some finite size effect. A marked difference occurs between the cases "." (dot) and "," (comma) on one hand and the others, ":", ";", "!", "?", i.e. colon, semi-colon, exclamation point, question mark, respectively. In the first group, the apparently best slope is rather close to 1/3, but is closer to 1/2 for the latest four cases, see examples in Fig. 2. It is obvious that in the latter cases a linear fit is even rather poor.

To find $\zeta = 1$ is usual, but low values like 0.50 and 0.33 are more rare. A value smaller than unity indicates a more homogeneous repartition of the variables (words, here). Whence a Gabaix [66] or Simon [67] model can be thought of to understand the values found here. E.g. Gabaix, studying city sizes, claims that two causes can lead to $\zeta \leq 1.0$, i.e. either:

Table 3
Values of parameters for the Z–M fit, Eq. (2), for various texts and sentences defined through various marks; the fit range is also indicated

	Text	A	C	ζ^*	Range	
					$\ldots \leq R \leq \ldots$	
Whole text	AWL_{eng}	1 177	0.17	1.15	2 ...	... 1000
	AWL_{esp}	962	0.28	1.01	2 ...	... 1000
	TLG_{eng}	1 098	0.13	1.21	2	... 1000
Chap. 1	AWL_{eng}	116	0.19	1.16	1...	...200
	AWL_{esp}	48	0.15	0.01	4 ...	...200
Chap. 2	AWL_{eng}	118	0.24	1.07	1 ...	...200
	AWL_{esp}	168	0.90	0.92	2 ...	...200
Dot.	AWL_{eng}	1 062	0.08	0.55	4 ...	...200
	AWL_{esp}	984	0.5	0.36	1 ...	...200
	TLG_{eng}	1 029	0.46	0.34	1 ...	...200
Comma,	AWL_{eng}	366	0.09	0.33	1 ...	...200
	AWL_{esp}	1 019	1.2	0.34	4 ...	...200
	TLG_{eng}	382	0.11	0.27	1 ...	...200
Semi-colon :	AWL_{eng}	4 650	0.06	1.15	2 ...	... 100
	AWL_{esp}	6 978	0.14	0.97	1 ...	... 100
	TLG_{eng}	13 128	0.02	4.73	1 ...	...60
Colon ;	AWL_{eng}	5 068	0.34	0.59	1 ...	... 100
	AWL_{esp}	5 645	0.3	0.6	1 ...	... 100
	TLG_{eng}	3 296	0.05	0.94	1 ...	... 100
Exclam. !	AWL_{eng}	2 756	0.03	1.48	4 ...	...200
	AWL_{esp}	3 630	0.02	1.91	3 ...	...200
	TLG_{eng}	4 777	0.39	0.61	1 ...	...200
Quest. ?	AWL_{eng}	3 283	0.01	3.61	4 ...	... 100
	AWL_{esp}	4 099	0.02	1.89	2 ...	... 100
	TLG_{eng}	5 504	0.13	0.84	1 ...	... 100

(1) the growth process deviates from Gibrat's law [68] which assumes that the mean growth rate is independent of the size, or
(2) the variance of the growth process is size-dependent.

Simon [67] considers that words not yet used are added at a constant rate, while words already used are inserted at a frequency depending of the previous number of occurrences; this leads to a Zipf exponent depending on the ratio between the rate of appearance of new words and the text length rate of increase. Thus one can agree that the sample size is relevant for explaining a $\zeta \leq 1$ value. In the present case, the found values correspond to the length of various sentences which are defined through various punctuation marks and from counting characters rather than words. Different ζ values can indeed be distinguished through the order of magnitude in the maximum rank. What is still surprising is why the longest sentences, defined through dots and commas, lead to smaller values than those for other punctuation marks, i.e. for less frequent sentences.

An alternate view can be taken through the Z–M analytical form, Eq. (2). Values of the parameters are given in Table 3 for various ranges R. It is fair to state that the parameters are NOT very robust with respect to the range. Values of ζ^* can be found close to the apparently best looking slope, ζ, but other values can be found as well. This is due to the strong influence of the low rank points. The paradoxical situation occurs when one remembers that the analytical form is supposed to be used in order to take into account the finite value at $R = 1$. However, the curvature for the (small) function words markedly influences the outcome. In order to illustrate the point, a brief example is given in the Appendix.

3.2. Grassberger-Procaccia plots: LTS analysis

The analysis of correlation integrals allow to estimate whether the number of degrees of freedom of a process is large or reasonably small. It seems that the usual goal is rather qualitative. However it pertains to the fundamental question on noisy signals, — is it noise or chaos? As explained here above the algorithm is based on the statistics of pairwise distances for an arbitrary choice of the delay time. Therefore the output of the method results in observing an evolution of correlations, i.e. in the knowledge on how often a point in "the" phase space is found near another, whence illustrating some dynamical features connecting local and global features.

The three sets of correlation integrals, calculated following the method here above described, are shown in Ref. [64]; they are similar to the example selected for Fig. 3. The slopes can be summarized through a graph relating r and n (Fig. 4). It is found that the attractor dimension n is not only smaller than the space dimension, as it should be [62], but also is a linear function on a log–log plot of the so called phase space dimension r, *for the three texts of interest*. Whatever the text, a remarkable power law is found, $r = n^{\lambda}$, with $\lambda = 0.79$, which does not indicate any saturation. Thus it seems interesting to examine the universality of such a law in further work, i.e., other authors and to find whether λ characterizes some author or some style of writing, or specific book natures.

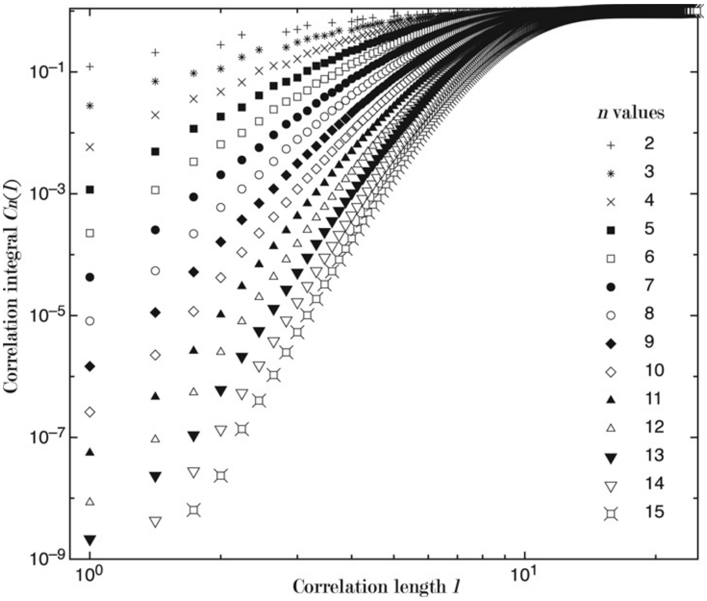

Fig. 3. Illustrative example: Grassberger–Procaccia (log–log) plot of the correlation integral $C_n(l)$ as a function of the correlation length l in phase spaces with different dimensions (n) for TLG_{eng}.

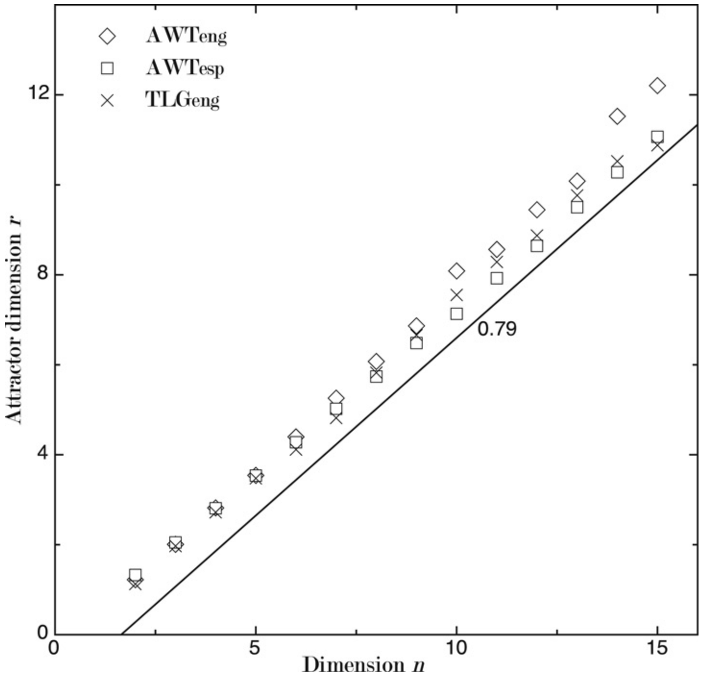

Fig. 4. Attractor dimension n as a function of the so called phase space dimension r for the three texts of interest. Notice that a linear relationship is found with a proportionality coefficient $\lambda = 0.79$ on a log–log plot, as for $r = n^{\lambda}$.

4. Conclusion

At first sight, a time series of a single variable appears to provide a limited amount of information. One usually thinks that such a series is restricted to a one-dimensional view of a system, which, in reality, contains a large number of independent variables. Nevertheless, one can map texts to univariate time series. Here FTS and LTS can be thought to originate from a dynamical process. A mere statistical analysis of the output, as done through a Zipf-like approach has been used. Power laws with different characteristic exponents for the ranking properties have been found. The Zipf exponent can take values about 1.0, 0.50 or 0.30, depending on how a sentence is defined. Some finite size effect cannot be disregarded. However this

Table 4
Effect of low ranking points on Z–M fit; parameter values and their corresponding error bar for AWL_{eng} sentences limited by “dots”

AWL_{eng}	Parameter	Value	Absolute error
Range: from 2 to 200	A	2104.5665	102.1277
	C	1.2151	0.2201
	ζ^*	0.3924	6.1668e−3
Range: from 2 to 200	A	1239.7700	18.5571
	C	0.1553	1.1255e−2
	ζ^*	0.4874	7.7509e−3
Range: from 3 to 200	A	1105.3531	10.7540
	C	9.4509e−2	4.7342e−3
	ζ^*	0.5334	6.9828e−3
Range: from 4 to 200	A	1061.8008	9.7680
	C	7.9410e−2	3.7663e−3
	ζ^*	0.5526	7.1248e−3

non-universality could be conjectured to be a measure of the author or book *style*, in view of the systematic set of results. One may immediately suggest to compare different authors known for different style schools together in further work. A more sophisticated approach through a Zipf–Mandelbrot law seems unreliable due to the (present lack) of discrimination between function and determining words, likely inducing breaks in the $f(R)$ plots. Something which has not been examined and is left for further studies is the distinction between oral-like and descriptive parts of a text and its translation.

Moreover, a time series is known [44,45,59] to bear the marks of all other variables participating in the dynamics of the system. Thus one claims to reconstruct the system phase space from such a series of one-dimensional observations. When applying the Grassberger–Proccacia (GP) method to a physics time series one wants to know whether the attractor is based on a finite set of variables. The lack of saturation found here through the law $r = n^{\lambda}$ for the size of the attractor indicates that the writing of a text by some creative author can be hardly reduced to a finite set of differential equations! Yet this simple analytical form suggests to examine whether λ characterizes an author style versatility or even creativity, and how robust its value can be, e.g. with regards to more modern authors.

Finally, as in Ref. [23] one can concur that the application of GP analysis indicates that linguistic signals may be considered as the manifestation of a complex system of high dimensionality, different from random signals or systems of low dimensionality such as the Earth climate or financial signals. I consider that this is mainly due to the fact that human languages must obey grammatical rules.

Last but not least as on comparing AWL_{eng}, AWL_{esp}, and TLG_{eng}, with both the “static” and “dynamic” methods, it seems that the texts are qualitatively similar, which indicates …the quality of the translator. In this spirit, it would be interesting to compare results originating from texts obtained through human and machine translations [69]. It is of huge interest to see whether a machine is *more flexible* with vocabulary and grammar than a human translator, — see also Ref. [39]!

Acknowledgements

The author would like to thank D. Stauffer, as well as E. Eulaers and M. Kefer for fruitful discussions on linguistic topics. This work has been partially supported by European Commission Project CREEN FP6-2003-NEST-Path-012864 (CREEN: Critical Events in Evolving Networks) and European Commission Project E2C2 FP6-2003-NEST-Path-012975 (E2C2: Extreme Events: Causes and Consequences).

I am very thankful to Jeremie Gillet to allow me to publish the results of his data analysis and for many comments.

Appendix

The Zipf–Mandebrot, Eq. (2), law is thought to be useful for better describing the ranking function $f(R)$, in particular in order to take into account the finite value of f at $R \simeq 1$. Yet from data presented in Table 4, it can be observed that the parameters, in particular $\zeta^*$are far from robust when the range of R slightly varies. For example ζ^* can vary from 0.84 to 3.61 when only the fit range is slightly changed, for sentences limited by question marks in the three original texts where one expects an exponent near 0.5. It appears that if one fits from $R = 1$ one finds $\zeta^* = 0.65$, but from $R = 2$, $\zeta^* = 1.68$, from $R = 3$, $\zeta^* = 2.68$, and from $R = 4$, $\zeta^* = 3.61$. This is “obviously” due to the curvature of the data at low R.

I have not found much discussion of the matter in the literature, maybe because either the case is not frequent, or not examined. See nevertheless [70] where it is suggested that ζ^* be interval dependent and increasing logarithmically with R. In the present case, it appears that one can consider the origin of the instability to reside in the “large” variations of $f(R)$ at small R. In fact the curvature of $f(R)$ changes from convex to concave at small R. This leads to an instability in the set of least mean square fits. This, in other words, is due to the number of regimes, changes in curvature, in the data. Powers [14] (and later others like [65]) had already noticed that one should distinguish between small (function) words and large (determining) words, and pointed to the break, or change in slope at finite R ($\sim$100). A recommendation is in order: a visual scan of the data should be made before attempting a fit with Eq. (2), in order to observe the number of regimes, or the number of crossover points, which might appear in the data. Such considerations would likely illuminate in the present context, the vocabulary quality level of an author.

References

[1] J.M. Klinkenberg, Des langues romanes, Duculot, Louvain-la-Neuve, 1994.
[2] N. Chomsky, Reflections on Language, Pantheon Books, New York, 1975.
[3] C. Schulze, D. Stauffer, Computer simulation of language competition by physicists, in: B.K. Chakrabarti, A. Chakraborti, A. Chatterjee (Eds.), Econophysics and Sociophysics: Trends and Perspectives, Wiley-VCH, Weinheim, 2006, pp. 311–318.
[4] V.M. de Oliveira, M.A.F. Gomes, I.R. Tsang, Physica A 361 (2006) 361.
[5] M. Nowak, D. Krakauer, Proc. Natl. Acad. Sci. USA 96 (1999) 8028.
[6] A. Baronchelli, L. DallAsta, A. Barrat, V. Loreto, Phys. Rev. E 73 (2006) 015102 R.
[7] D. Benedetto, E. Caglioti, V. Loreto, Phys. Rev. Lett. 88 (2002) 048702.
[8] L. Steels, Artif. Life 2 (1995) 319;
L. Steels, Evolution Commun. 1 (1997) 1.
[9] R. Lambiotte, M. Ausloos, M. Thelwall, J. Informetrics 1 (2007) 277.
[10] G.K. Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley, Cambridge, 1949.
[11] G.K. Zipf, The Psycho-Biology of Language, An Introduction to Dynamic Philology, Houghton Mifflin, Boston, 1935.
[12] G.U. Yule, Statistical Study of Literary Vocabulary, Cambridge Univ. Press, 1944.
[13] C.T. Meadow, J. Wang, M. Stamboulie, J. Inform. Sci. 19 (1993) 247.
[14] D.M.W. Powers, Applications and explanations of Zipf's laws, in: D.M.W. Powers (Ed.), New Methods in Language Processing and Computational natural Language Learning, ACL, 1998, pp. 151–160. NEMLAP3/CONLL98.
[15] B.J. West, W. Deering, The Lure of Modern Science: Fractal Thinking, World Sci, River Edge, NJ, 1995.
[16] R. Ferrer i Cancho, Eur. J. Phys. B 44 (2005) 249.
[17] K. Kawamura, N. Hatano, Models of Universal Power-Law Distributions, cond-mat/0303331.
[18] K. Kawamura, N. Hatano, J. Phys. Soc. Japn 71 (2002) 1211; J. Phys. Soc. Japn 72 (2003) 1594 (erratum).
[19] A.F. Gelbukh, G. Sidorov, Zipf and Heaps Laws' coefficients depend on language, in: Proc. Second International Conference on Computational Linguistics and Intelligent Text Processing, vol. 24, 2001, pp. 332–335.
[20] E. Métois, Musical sound information: Musical gestures and embedding synthesis, Ph.D. Thesis, MIT, (Chapter 4) (unpublished).
[21] D.H. Zanette, M.A. Montemurro, J. Quant. Linguist. 12 (2005) 29.
[22] M.A. Montemurro, D.H. Zanette, Glottometrics 4 (2002) 86.
[23] K. Kosmidis, A. Kalampokis, P. Argyrakis, Physica A 370 (2006) 808.
[24] N. Hatzigeorgiu, G. Mikros, G. Carayannis, J. Quant. Linguist. 8 (2001) 175.
[25] A. Schenkel, J. Zhang, Y.-C. Zhang, Fractals 1 (1993) 47.
[26] M. Amit, Y. Shemerler, E. Eisenberg, M. Abraham, N. Shnerb, Fractals 2 (1994) 7.
[27] M.A. Montemurro, Physica A 300 (2001) 567.
[28] M. Montemurro, P. Pury, Fractals 10 (2002) 451.
[29] L.Q. Ha, E.I. Sicilia-Garcia, J. Ming, F.J. Smith, J. Comput. Linguist. Chin. Lang. Process. 8 (2003) 77.
[30] W. Ebeling, A. Neiman, Physica A 215 (1995) 233.
[31] M. Boulton, Zamenhof, Creator of Esperanto, Routledge, Kegan & Paul, London, 1960.
[32] B. Manaris, L. Pellicoro, G. Pothering, H. Hodges, Investigating Esperanto's statistical proportions relative to other languages using neural networks and Zipf's law, in: Proc. 24th IASTED Int. Conf. on Artificial Intelligence and Applications, ACTA Press, Anaheim, CA, 2006, pp. 102–108.
[33] http://www.languagemonitor.com/.
[34] http://www.en.wikipedia.org/wiki/Plena_-Ilustrita_-Vortaro_-de_-Esperanto.
[35] Ch. Vander invented a constructed language, Kobaian, in which most lyrics of his progressive rock band, Magma, are sung; see http://www.en.wikipedia.org/wiki/Magma_-(band).
[36] Urban Trad music group participated in the Eurovision Song Contest 2003, where they ended second with the song Sanomi, a modern folk song with vocals in an imaginary language; cf. http://www.en.wikipedia.org/wiki/Urban_-Trad.
[37] When finalizing the writing of this paper I became aware of Refs. [14,38,39], discussing translated texts. In particular, Powers [14] considered the Bible in four languages, as a sequence of stories, just like Alice in Wonderland.
[38] D.R. Amancio, L. Antiqueira, T.A.S. Pardo, L. da, F. Costa, O.N. Oliveira Jr., M.G.V. Nunes, Int. J. Mod. Phys. C 19 (2008) 583.
[39] I. Kanter, H. Kfir, B. Malkiel, M. Shlesinger, J. Quant. Linguist. 13 (2006) 35.
[40] National Clearinghouse for Machine Readable Texts, Project Gutenberg, 2005. http://www.gutenberg.org.
[41] L.W. Carroll, Alice's Adventures in Wonderland, Macmillan, 1865, see http://www.gutenberg.org/etext/11.
[42] L.W. Carroll, Through the Looking Glass and What Alice Found There, Macmillan, 1871, see http://www.gutenberg.org/etext/12.
[43] Esperanto texts can be found in D. Harlow, Literaturo, en la reto, en Esperanto, (2005), accessed Sep. 28, 2005. see http://www.donh.best.vwh.net/Esperanto/Literaturo/literaturo.html; http://www.meeuw.org/bibliografio/8/verloren.html; http://www.esperanto.net/literaturo/; http://www.esperantujo.org/eLibrejo/.
[44] P. Grassberger, I. Procaccia, Phys. Rev. Lett. 50 (1983) 346.
[45] P. Grassberger, I. Procaccia, Physica D 9 (1983) 189.
[46] W. Li, see http://www.linkage.rockefeller.edu/wli/zipf/.
[47] S. Abe, N. Suzuki, Physica A 350 (2005) 588.
[48] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.K. Peng, M. Simons, H.E. Stanley, Phys. Rev. Lett. 73 (1994) 3169.
[49] Ph. Bronlet, M. Ausloos, Int. J. Mod. Phys. C. 14 (2003) 351.
[50] R. Rousseau, A weak goodness-of-fit test for rank-frequency distributions, in: C. Macias-Chapula, (Ed.), Proceedings of the Seventh Conference of the International Society for Scientometrics and Informetrics, Universidad de Colima Mexico, 1999, pp. 421–430.
[51] B. Mandelbrot, Information theory and psycholinguistics: A theory of words frequencies, in: P. Lazafeld, N. Henry (Eds.), Readings in Mathematical Social Science, MIT Press, Cambridge, MA, 1966.
[52] B.B. Mandelbrot, An informational theory of the statistical structure of languages, in: W. Jackson (Ed.), Communication Theory, Betterworth, 1953, pp. 486–502. xxx.
[53] B.B. Mandelbrot, Simple games of strategy occurring in communication through natural languages, in: Symposium on Statistical Methods in Communication Engineering, Transactions of IRE 3 (1954) 124.
[54] W. Li, IEEE Trans. Inform. Theory 38 (1992) 1842.
[55] K. Kosmidis, J.M. Halley, P. Argyrakis, Physica A 353 (2005) 595.
[56] B. Vilenski, Physica A 231 (1996) 705.
[57] Ch. Karakotsou, A.N. Anagnostopoulos, Physica D 93 (1996) 157.
[58] Ch.L. Koliopanos, I.M. Kyprianidis, I.N. Stouboulos, A.N. Anagnostopoulos, L. Magafas, Chaos Solitons Fractals 16 (2003) 173.
[59] G. Nicolis, I. Prigogine, Exploring Complexity, Freeman, New York, 1989.
[60] E.J. Kostelich, H.L. Swinney, Phys. Script. 40 (1989) 436.

[61] J. Theiler, J. Opt. Soc. Am. 7 (1990) 1055.
[62] F. Takens, On the Numerical Determination of the Dimension of an Attractor, in: Lect. Notes Math., vol. 1125, Springer, Berlin, 1985, pp. 99–106.
[63] J.P. Zbilut, C.L. Webber, Phys. Lett. A 237 (1998) 131.
[64] A more complete set of figures can be found in M. Ausloos, Equilibrium (Zipf) and Dynamic (Grassberger-Procaccia) method based analyses of human texts. A comparison of natural (English) and artificial (Esperanto) languages, http://arxiv.org/pdf/0802.4215v1.
[65] R. Ferrer i Cancho, R.V. Solé, J. Quant. Ling. 8 (2002) 165.
[66] X. Gabaix, Quat. J. Econom. 114 (1999) 739.
[67] H.A. Simon, Biometrika 42 (1955) 425.
[68] R. Gibrat, Les inegalités économiques, Librairie du Recueil, Paris, 1931.
[69] A. Koutsoudas, Language 33 (1957) 544.
[70] I. Kanter, D.A. Kessler, Phys. Rev. Lett. 74 (1995) 4559.