entropy

MDPI

Article

Evaluating the Irregularity of Natural Languages

Candelario Hernández-Gómez [1], Rogelio Basurto-Flores [2], Bibiana Obregón-Quintana [3] and Lev Guzmán-Vargas [2,*]

[1] Departamento de Física, Escuela Superior de Física y Matemáticas, Instituto Politécnico Nacional, Ciudad de México 07738, Mexico; hernandezgomez2010@gmail.com

[2] Unidad Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, Instituto Politécnico Nacional, Ciudad de México 07340, Mexico; rogelio.basurto@gmail.com

[3] Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad Universitaria, Ciudad de México 04510, Mexico; b.obregon.q@gmail.com

* Correspondence: lguzmanv@ipn.mx

Received: 9 August 2017; Accepted: 25 September 2017; Published: 29 September 2017

Abstract: In the present work, we quantify the irregularity of different European languages belonging to four linguistic families (Romance, Germanic, Uralic and Slavic) and an artificial language (Esperanto). We modified a well-known method to calculate the approximate and sample entropy of written texts. We find differences in the degree of irregularity between the families and our method, which is based on the search of regularities in a sequence of symbols, and consistently distinguishes between natural and synthetic randomized texts. Moreover, we extended our study to the case where multiple scales are accounted for, such as the multiscale entropy analysis. Our results revealed that real texts have non-trivial structure compared to the ones obtained from randomization procedures.

Keywords: approximate entropy of texts; sample entropy; text irregularity; symbol sequences

1. Introduction

Diverse studies have reported spatio-temporal organization properties in natural languages. Two representative findings of universal features of natural language are the Zipf and Heaps laws, which are based on word frequency and number of different words, respectively [1–4]. From a more basic perspective, human language can also be considered as a sequence of symbols which contains information encoded in the patterns (words) needed to communicate. For instance, the frequency rate of appearance of the symbols is different for every language, and so are the declension and verbal conjugation rules. There are also restrictions in the order of appearance of bigrams, trigrams and, in general, n-grams; for example, in English and Spanish the letter "q" is always followed by "u". The way these restrictions and other factors modulate the structure and randomness of the language can be potentially evaluated by means of concepts like entropy, as proposed by Shannon [5,6]. The use of entropy-related algorithms to estimate orderliness in natural language have revealed that language is not regular nor random, but the direct quantification of the presence of randomness is not an easy task. Diverse studies have used the concept of entropy by means of a n-gram analysis [7], a binary simplification [8], nonparametric entropy estimation [9], mutual information of letters [10], information-based energy [11], complexity quantification [12] and entropy-word approach [13]. However, entropy-information measures based on regularity of pattern statistics has not been widely employed to evaluate the "complexity" of natural language. A straightforward way to quantify the propinquity of two words is to count the number of letters that they have in common; the words coming from the same root, diminutives, augmentatives or the "functional shift" of certain words are good examples of these similarities. In the context of dynamical systems, there are well known methods to measure the repetition pattern in a time series: the approximate entropy ($ApEn$) and its derivatives [14,15]. The $ApEn$ quantifies the regularity in a time series, a lower value of $ApEn$ indicates a more regular behavior whereas a high value is assigned to more irregular sequences. This method has been successfully applied to analyze time series from several sources [16–20]. Here we adopt a similar approach based on the $ApEn$ algorithm in order to evaluate the levels of complexity in four language families (Romance, Germanic, Slavic and Uralic). Our goal is to determine the dominance of regularities in written texts, which are considered as finite time series. Our method was applied to several texts from different languages. The results reveal that, for texts belonging to the same family, it is observed that the $ApEn$ decreases as the length of the word pattern increases in similar fashion. Moreover, we also extend our study to evaluate the multiscale behavior of entropy for the assessment of regularities based on different scales as it was suggested by Costa et al. [21]. Additionally, we also apply our methodology to two synthetic sequences, which are the randomized versions of the original text and a text written in Esperanto. We found significant differences between real and synthetic texts, observing a higher complexity for the real sequences compared to the randomized ones through different scales. The paper is organized as follows: First, we present the methodology used throughout the article, including the modified method to calculate the $ApEn$ for the cases of sequences of symbols. Next, the main results of the study are described; and finally, we provide some final remarks.

2. Approximate Entropy of a Text

Within the context of information theory, the entropy of a sequence of symbols (from an alphabet with L elements) is given in terms of the so-called Shannon entropy $H_S = -\sum_{j=1}^{L} p_j \log p_j$, with p_j the probability of the symbol j. The Shannon entropy measures the average uncertainty of a discrete variable and represents the average information content [6]. For sequences composed of blocks with n symbols, the entropy $H_n = -\sum_j p_j^{(n)} \log p_j^{(n)}$ measures the uncertainty assigned to a word of length n [22,23]. The difference entropy $h_n = H_{n+1} - H_n$ represents the uncertainty related to the appearance of the $n+1$ symbol given that the n preceding symbols are known [22]. For dynamical systems, the estimation of the mean rate of creation of information is given by the Kolmogorov–Sinai (KS) entropy and KS measures the unpredictability of systems changing with time [24]. However, numerical calculations of KS requires very large sequences, therefore it is not practical to apply to real sequences. In order to overcome this limitation, Grassberger et al. [25] proposed the K_2 entropy to evaluate the dimensionality of chaotic systems as a lower bound of the KS entropy. Later, as an extension of the K_2 entropy, Pincus [14] introduced the Approximate Entropy ($ApEn$) to evaluate the regularity in a given time series. The $ApEn$ provides a direct measure of the degree of irregularity or randomness in a time series and, in the context of physiological signals, as a measure of system complexity: smaller values indicate greater regularity, and greater values convey more disorder or randomness [14,17]. Here we introduce a modified $ApEn$ algorithm for the regularity analysis of a written text. Our method considers a similar procedure as the $ApEn$ proposed by Pincus [14] and it can be summarized as follows: for a given text, $\{s(1), s(2), s(3), ..., s(N)\}$ of N elements, where an element can be a letter or symbol (including the space), we define a set of patterns of length m, $S_m(i)$ for $i \in [1, N-m+1]$, where $S_m(i) = \{s(i+k) | 0 \leq k \leq m-1\}$ is the pattern of m elements or symbols, from $s(i)$ to $s(i+m-1)$. Next, we look for matches occurring between two patterns if the "distance" is smaller than a given value. We impose a restriction to the "distance" between two such patterns, i.e., we set a number r representing the maximum number of positions at which the corresponding symbols are different. This distance is known as the Hamming distance [26]. Next, we calculate the number n_i^m of patterns $S_m(j)$ with $j \leq N-m+1$ such that $h(S_m(i), S_m(j)) \leq r$, with $h(S_m(i), S_m(j))$ the Hamming distance. Then, the quantity $C_i^m(r) = \frac{n_i^m}{N-m+1}$ is defined, representing the probability of having patterns within the distance r from the template pattern $S_m(i)$.

Following Pincus [14] we define the Approximate Entropy in the case of texts as,

$$ApEn(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r), \tag{1}$$

where $\Phi^m(r)$ and $\Phi^{m+1}(r)$ are given by $\Phi^m(r) = \frac{1}{N-m+1}\sum_{i=1}^{N-m+1}\ln C_i^m(r)$ and $\Phi^{m+1}(r) = \frac{1}{N-m}\sum_{i=1}^{N-m}\ln C_i^{m+1}(r)$, respectively. As in the context of time series, the statistic represented by $ApEn$ quantifies the degree of regularity/irregularity in a given text, and it is conceived as approximately equal to the negative average natural logarithm of the conditional probability that two patterns that are similar for m symbols remain similar for $m+1$ elements [14]. Although $ApEn$ is very useful for distinguishing a variety of deterministic/stochastic processes, it has been reported that there is a bias in $ApEn$ because the method counts each pattern as matching itself. The existence of this bias, under particular circunstances, causes $ApEn$ to substimate or to provide a faulty value for a given time series. Therefore, the development of an alternative method was desirable to overcome the limitations of $ApEn$. On the basis of K_2 and $ApEn$ algorithms, Richman and Moorman [15] introduced the so-called sample entropy ($SampEn$) to reduce the bias in $ApEn$. One of the advantages of $SampEn$ is that it does not count self-matches and is not based on a template-wise approach. Discounting the self-matches is justified since the entropy is conceived as a measure of the rate of information production; then, self-matches do not add new information [15]. Following the definition of Richman and Moorman [15], we can also define the $SampEn(m,r,N)$ in the case of texts as,

$$SampEn(m,r,N) = -\ln\frac{U^{m+1}}{U^m}, \tag{2}$$

where U^{m+1} and U^m are the probabilities that two patterns will match (with a tolerance of r) for $m+1$ and m symbols, respectively [27]. As in the case of $ApEn$, $SampEn$ is conceived as the negative natural logarithm of the conditional probability that two sequences similar for m points remain similar at the next point, with a tolerance of r, without counting the self-matches; and a lower value of $SampEn$ indicates a more regular behavior of a symbol sequence whereas high values are assigned to more irregular sequences. We remark that both $ApEn$ and $SampEn$ represent family statistics that depend on the sequence length N, the tolerance parameter r and the pattern length m.

3. Results and Discussion

Prior to the description of our results, we briefly explain the main steps of our method for a simple case of a very short text. Lets consider the beginning of the famously acknowledged Hamlet's soliloquy: *To-be-or-not-to-be*. The length of the sentence is 18, and the average length of words in this sentence is 2, i.e., approximately every three symbols the space mark repeats, then a natural value for m is 3, and we set the tolerance value $r = 1$ (33% of the pattern length). Starting with the first letter from the left, the 16 subseries we can built are $S_1 = \{$*To-*$\}$, $S_2 = \{$*o-b*$\}$, ..., $S_{16} = \{$*-be*$\}$. After performing all the procedure described in the previous section, we find that, for the Hamlet's soliloquy beginning, the statistics Equation (1) results in $ApEn(3,1,18) = 0.215$. This is a relatively intermediate value, which indicates that the sentence is moderately predictable compared to the case where the position of symbols was randomized ($ApEn_{rand} = 0.435$ in average for five independent realizations).

First, we analyze literary texts from each of the 14 languages which are described in Table 1 for an extended dataset, which includes two more books of each language, please refer to the supporting information online at [28]. The texts were downloaded from the website of the Gutenberg Project (http://www.gutenberg.org). In order to avoid finite size effects and to validate our method for relatively short sequences, we restrict ourselves to segments with 5000 symbols and repeat the calculations for 10 segments this length [29]. In our case we have kept the punctuation marks and the space mark as symbols. In what follows, we will only refer to $ApEn$ values since we obtain the same qualitative results using either $SampEn$ or $ApEn$ algorithms and particular differences will be discussed elsewhere.

Table 1. Books written in different languages considered in our study. For each book, we also include the linguistic family, the language, the number of symbols in the alphabet (L), the mean $ApEn$ values for $m = 5$ and $m = 6$ with $r = 2$, the total number of words N and the total number of different words M. The Esperanto text $(L = 28)$ used in the analysis is La Batalo del Vivo, of Charles Dickens. For an extended dataset see supporting information online at [28].

European Macrofamily							
Family	**Language**	L	**Title (Author)**	$ApEn_{m=5}$ **(S.D.)**	$ApEn_{m=6}$ **(S.D.)**	N	M
Romance	Latin	23	Commentaries on the Gallic War (Julius Caesar)	1.221 (0.028)	0.8558 (0.045)	51,643	11,688
	Spanish	27	His only son (Leopoldo Alas)	1.1321 (0.028)	0.9226 (0.032)	89,372	12,651
	Italian	21	Memories of Paris (Edmondo de Amicis)	1.216 (0.038)	0.894 (0.046)	54,509	11,073
	French	26	Zadig (Voltaire)	1.132 (0.045)	0.903 (0.378)	28,099	4676
Germanic	English	26	Moby Dick (Herman Melville)	1.079 (0.044)	0.762 (0.030)	214,903	18,772
	German	30	The Golden Pot (E.T. Hoffman)	1.201 (0.032)	0.748 (0.033)	29,252	6089
	Swedish	29	Sara Videbeck and the Chapel (C.J.L. Almqvist)	1.162 (0.034)	0.748 (0.033)	36,899	6702
	Dutch	28	Sense and sensibility (Jane Austen)	1.165 (0.033)	0.862 (0.048)	125,965	9433
Slavic	Rusian	33	Anna Karenina (L. Tolstoy)	1.105 (0.087)	0.611 (0.056)	271,387	35,308
	Polish	32	One month of prose works (Jan Sten)	1.149 (0.049)	0.754 (0.070)	31,452	10,233
	Serbian	30	Stone age (Jovan Zvjovic)	1.133 (0.069)	0.710 (0.097)	37,065	9465
	Czech	42	The double (F.Dostoyevsky)	1.047 (0.182)	0.569 (0.158)	120,768	21,078
Uralic	Finnish	29	Erkki Ollikainen (J.O. Åberg)	1.264 (0.029)	0.898 (0.040)	25,185	8855
	Hungarian	44	A három galamb (Kadar Lehel)	1.1952 (0.022)	0.7912 (0.040)	30,400	9800

Figure 1 shows the calculations of the average $ApEn$ for several values of m and a fixed value $r = 2$. We notice that for $m = 6$, the $ApEn$ values obtained for each language tend to be close when they are grouped according to the family to which they belong, allowing a comparison between families (see Table 1). For this m value, the Romance family exhibits the highest value of $ApEn$, followed by the Uralic, Germanic and Slavic families, indicating that different levels of regularity/irregularity are observed in the analyzed family languages (Figure 1a–d). It is also worth to mention that languages that belong to the same family display a similar profile as the pattern length increases, while the value of entropy for Romance, Slavic and Uralic families almost monotonically decreases; for the Germanic one the entropy exhibits a small change between $m = 6$ and $m = 7$, revealing that the level of irregularity remains approximately constant for these pattern length scales, which roughly corresponds to the mean word length of these languages [7]. Notably, all the Romance languages are much more overlapped compared to ApEn curves for the other families.

To further compare the $ApEn$ profiles in texts, we also studied two artificial cases: Esperanto and randomized versions of the original sequences. Invented languages like Esperanto are attempts to simplify natural languages by suppressing, for example, irregular verbs and including words from different languages to make it universal. For the randomized version, we consider a text which is randomly generated with identical symbol and space probabilities as observed in a real text and ten independent realizations were constructed. The results of $ApEn$-values for the randomized versions are shown in Figure 1f. In Figure 1e we also show the behavior of entropy in terms of m for Esperanto. For $r = 2$, we observe that at $m = 3$, the entropy value is close to the values observed in the majority of natural languages, and a rapid decay is observed between $m = 4$ and $m = 5$, being this decline much faster than the one observed in real texts (see Figure 1a–d). We note that for values between $m = 3$ and $m = 4$, a higher value of $ApEn$ is observed for random texts than for real texts, and then the entropy decreases dramatically for larger values of the pattern length. We remark that for short length patterns the $ApEn$ is high due to the fact that the frequency of m and $m + 1$ length patterns is quite different, indicating a high irregularity in the text, as expected for random sequences. When the entropy values (corresponding to pattern lengths 3–10) from the different languages were pairwise compared with their corresponding random version, we found significant differences in almost all cases (p-value < 0.05 by Student's test, see Table 2 for details).

In order to further characterize the effects of the parameters m and r on the entropy values, in Figure 2 we show the calculations of $ApEn$ for 36 pairs of values of the parameter r and the pattern length m. Recall that the r value represents the similarity criterion based on the Hamming distance, i.e., the number of positions at which the corresponding symbols may differ. Thus, r takes values between 1 and $m-1$. As shown in Figure 2, we find that in most cases the entropy increases as the parameter r increases for a fixed value of the pattern length m, whereas for a fixed r the entropy value tends to decrease as m increases. Note that, for Germanic languages this general behavior is not observed as m increases (Figure 2(b1)). As a general remark of the dependence of entropy on parameters m and r, we notice that an acceptable value of r is given by the level of discrepancies between the two patterns (a factor of the pattern length), since for larger m values and small values of r, a higher concordance is required every time, i.e., almost a perfect match, and larger sequences are required to get a reliable statistic.

Finally, to compare the behavior of the entropy values, we applied the Fisher's linear discriminant [30] to the data showed in Figure 1a–d. This technique is very useful to determine if the $ApEn$ profiles could potentially classify languages into the Romance, Germanic, Slavic and Uralic families. Results for the 14 languages are presented in Figure 3. For this analysis we considered the $ApEn$ values (corresponding to pattern lengths 3–10) from ten segments of 5000 symbols for each language. Then, the data were projected down to a two-dimensional scatter plot presented in Figure 3. We observe a separation between clusters formed by languages that belong to the same linguistic family.

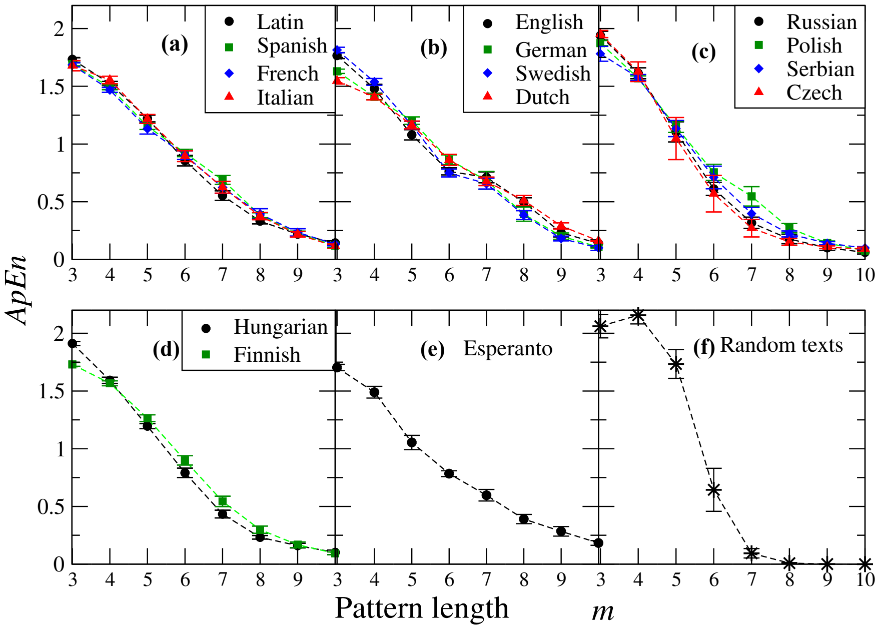

Figure 1. Approximate entropy ($ApEn$) as a function of the pattern length m for four families of European languages. In all cases we set $r=2$. Here symbols represent the mean value of the entropy measure for 10 segments, each with 5000 symbols and the error bars represent the standard deviation. The four families considered here are: (**a**) Romance (Latin, Spanish, Italian and French); (**b**) Germanic (English, German, Swedish and Dutch); (**c**) Slavic (Russian, Polish, Serbian and Czech); and (**d**) Uralic (Finnish and Hungarian). We also show the cases of (**e**) Esperanto and (**f**) random versions of the cases in panels (**a**–**d**). For natural languages, we observe similarities in the decay profile of the entropy between languages which belong to the same family. We notice that for the Germanic family, the entropy measure remains almost constant for pattern lengths between $m=6$ and $m=7$. Esperanto shows a fast decay between $m=4$ and $m=5$, while random texts present a high value of entropy for $m=3$ and $m=4$ with abrupt decay from values greater than 4. For results of the extended dataset see supporting information online at [28].

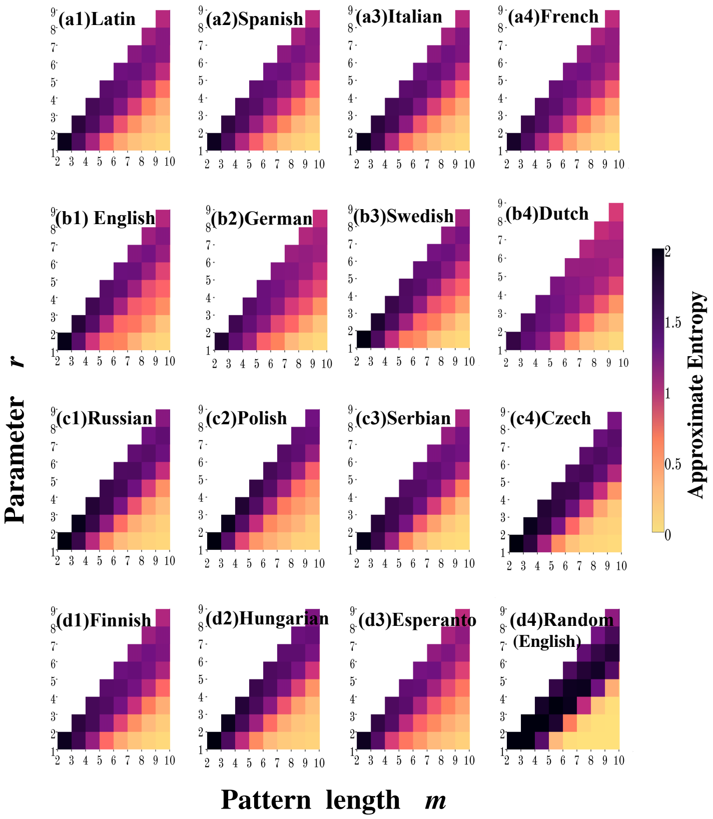

Figure 2. Behavior of $ApEn$ as a function of m and r for (**a1**) Latin, (**a2**) Spanish, (**a3**) Italian, (**a4**) French; (**b1**) English, (**b2**) German, (**b3**) Swedish, (**b4**) Ducth; (**c1**) Russian, (**c2**) Polish, (**c3**) Serbian, (**c4**) Czech; (**d1**) Finnish, (**d2**) Hungarian, (**d3**) Esperanto and (**d4**) randomized English text.

Table 2. Results of the application of the t-Student's test to $ApEn$ values of original texts vs. random versions. We only show the values of p for $m = 4, 5, 6, 7$. We observe that for most of the cases $p << 0.05$, indicating significant differences between original vs. random data, except the case of Dutch for $m = 6$ (highlighted in bold).

Language	***p*-Value**			
	***m* = 4**	***m* = 5**	***m* = 6**	***m* = 7**
Latin	5.468×10^{-20}	1.036×10^{-16}	0.001	7.853×10^{-12}
Italian	6.629×10^{-15}	7.662×10^{-16}	2.950×10^{-6}	2.763×10^{-11}
French	4.331×10^{-22}	1.197×10^{-13}	1.957×10^{-8}	7.586×10^{-11}
Spanish	1.048×10^{-22}	4.911×10^{-19}	1.296×10^{-9}	3.790×10^{-13}
English	2.083×10^{-14}	1.649×10^{-12}	2.589×10^{-5}	4.630×10^{-12}
German	1.029×10^{-19}	1.469×10^{-14}	1.880×10^{-5}	1.316×10^{-9}
Swedish	5.874×10^{-21}	3.206×10^{-18}	6.945×10^{-7}	7.534×10^{-12}
Dutch	2.853×10^{-22}	8.664×10^{-16}	**0.610**	4.297×10^{-14}
Russian	6.486×10^{-14}	5.428×10^{-10}	4.621×10^{-6}	1.963×10^{-8}
Polish	2.164×10^{-23}	2.722×10^{-13}	6.902×10^{-10}	1.101×10^{-8}
Serbian	6.955×10^{-14}	2.669×10^{-10}	0.001	1.309×10^{-8}
Czech	6.131×10^{-11}	6.753×10^{-6}	0.001	1.608×10^{-6}
Hungarian	1.531×10^{-19}	6.322×10^{-19}	1.085×10^{-13}	1.569×10^{-11}
Finnish	1.813×10^{-20}	3.814×10^{-15}	2.075×10^{-9}	6.239×10^{-11}
Esperanto	3.463×10^{-13}	6.042×10^{-12}	0.035	2.047×10^{-10}

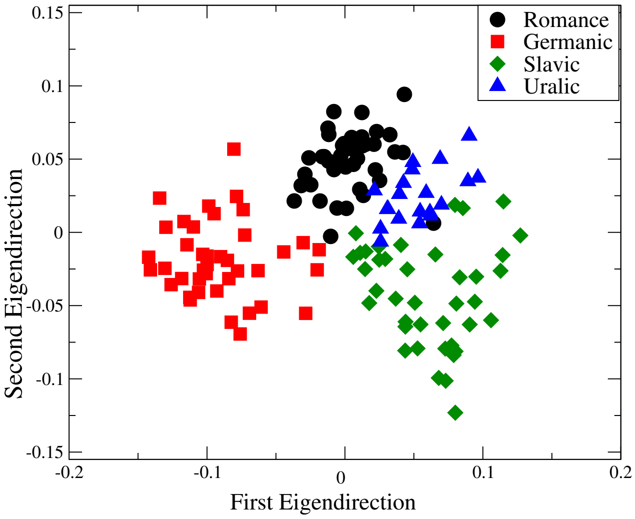

Figure 3. Results of the application of a linear classification analysis to data derived from four family languages. Here we show the projection of $ApEn$ values from patterns lengths between $m = 3$ and $m = 10$ (see Figure 1). For each m-value and for each language, we considered ten segments with length 5000 to obtain ten $ApEn$ values. Next, languages were labeled in classes according to the linguistic family to which they belong (Romance, Germanic, Slavic, Uralic). The eight dimensional vectors comprising the eight $ApEn$ (pattern lengths 3–10) values are used to create the two-dimensional projection. We observe that the families are segregated. For results of the extended dataset see supporting information online at [28].

Multiscale Entropy Analysis of Texts

In the context of biological signals, Costa et al. [21] introduced the multiscale entropy analysis (MSE) to evaluate the relative complexity of time series across multiple scales. This method was introduced to give an explanation to the fact that, in the context of biological signals, single-scale entropy methods (such as $ApEn$) assign higher values to random sequences from certain pathologic conditions whereas an intermediate value is assigned to signals from healthy systems [17,21]. It has been argued that these results may lead to erroneous conclusions about the level of complexity displayed by these systems. Here we adopt a similar approach with the idea of evaluating the complexity of written texts by accounting multiple time scales. We explain the main steps of the modified MSE for the analysis of texts. Given the original sequence $\{s(1), s(2), s(3), ..., s(N)\}$, a coarse-graining process is applied. A scale factor τ is considered to generate new sequences with elements formed by repeated concatenation of symbols from non-overlapping segments of length τ. Thus, the coarse-graining sequences for a scale factor τ are given by $y^{\tau}_{k,j} = s_{(k-1)\tau+j} \cdots s_{k\tau+j-1}$, with $1 \leq k \leq N/\tau, 1 \leq j \leq \tau$ and the dots denote concatenation. We observe that for $\tau = 1$, the original sequence is recovered, whereas for $\tau > 1$ the length of the new sequences is reduced to N/τ. We note that for each scale factor τ, there are τ coarse-grained sequences derived from the original one, as it was recently pointed out in the composite MSE [31,32]. Next, to complete the MSE steps the $ApEn$ algorithm is applied to the sequence $y^{\tau}_{k,j}$ for each scale to evaluate the regularity/irregularity in the new block-sequences. In order to improve the statistics, the entropy was calculated for all the jth coarse-grained time series for a given τ and the MSE value is given in terms of the average value of the entropies. Finally, the entropy value is plotted against the scale factor. A very simple example of the coarse-graining procedure can be illustrated for the Hamlet's soliloquy: *"To-be-or-not-to-be"*. For $\tau = 2$ we obtain $y^2_{1,1} = \{To\}, y^2_{2,1} = \{\text{-}b\}, ..., y^2_{9,1} = \{be\}$ and $y^2_{1,2} = \{o\text{-}\}, y^2_{2,2} = \{be\}, ..., y^2_{8,2} = \{\text{-}b\}$. We note that these new sequences have components formed by two-letter blocks which are the input for the modified $ApEn$ algorithm described in the previous Section. In practice, each new τ-block component is assumed as a single character for calculation purposes. Figure 4 presents the results of the MSE analysis for the real and synthetic texts described in the previous Section. The value of the entropy for scales one and two is higher for random texts than for real ones (Figure 4a–d). It is noticeable that for texts from natural language, as the scale factor increases, the entropy value decreases moderately compared to the rapid decreasing observed for synthetic random data such that for scales larger than 3, the entropy values for random sequences are smaller than the ones from original texts. Similar conclusions were obtained for Esperanto and its random version (data not shown). As it has been identified when MSE was applied to biological signals, it is observed signals that exhibit long-range correlated behavior are more complex than the uncorrelated ones. When applied to natural language, our results show that the temporal organization of natural languages (with some differences between them) exhibits more complex structure than the sequences constructed by randomizations. These results are also concordant with previous studies, which report the presence of long-range correlations in written texts [33,34].

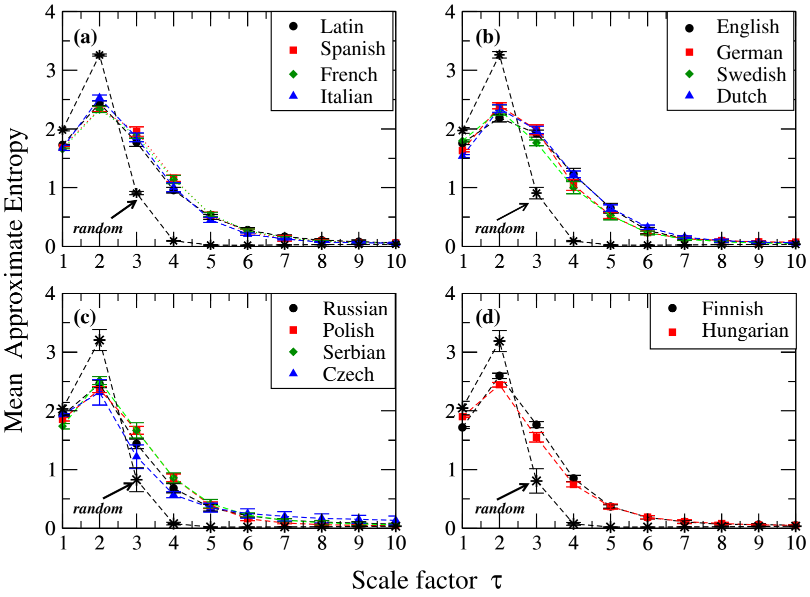

Figure 4. Multiscale entropy analysis (MSE) for 14 natural languages from 4 European families and their corresponding randomized sequences. Symbols represent the mean value of the *ApEn* for 10 segments and the error bars the standard deviation. The length of each segment is 5000 data elements and we used the values $m = 3$ and $r = 2$. Each panel shows the results for the (**a**) Romance; (**b**) Germanic; (**c**) Slavic and (**d**) Uralic families. Note that for scales one and two, randomized versions exhibit higher values of entropy than the real ones, while for scales bigger than three original sequences have more complexity vs random versions. We also note that random data from the Slavic and Uralic families remain very close to the entropy values of real texts compared to what happens to the Romance and Germanic cases, where a clear separation is observed for scale factors smaller than six.

4. Conclusions

We have presented a modified version of the approximate entropy method which is suitable for the evaluation of irregularities on multiple temporal scales in written texts from natural languages. First, we described the modified *ApEn* and *SampEn* methodologies by considering repetition of patterns subjected to a threshold Hamming distance. This entropy-based statistic (*ApEn*) was defined as approximately equal to the negative average natural logarithm of the conditional probability that two symbols-patterns that are similar for one length m remain similar when the length m is increased in one element [14]. We applied this algorithm to different natural languages which belong to four families. Our results showed that natural languages are neither regular nor random but exhibiting different levels of irregularity which are similar for languages belonging to the same family. The application of the Fisher linear discriminant analysis to the *ApEn*-values, revealed that the four language families are segregated. Besides, the modified MSE method was applied to compare the multiscale features in the same real texts as well as in randomized versions of themselves. We found that real sequences exhibit a non trivial structure compared to texts obtained from randomizations, i.e., natural languages have more complex structure observed across different local scales compared to sequences with arbitrary order. Finally, we point out that additional studies are needed to fully characterize natural language predictability as well as to consider corrections in the calculations of multiscale entropy values [32].

Acknowledgments: We thank F. Angulo, D. Aguilar, L. Arias, T. Contreras, I. Fernández, J. Fonseca, R. Hernández, F. Reyes, I. Reyes, N. Sánchez and J.C. López Vieyra for fruitful comments and suggestions. This work was partially supported by COFAA-IPN, EDI-IPN, SNI-Conacyt, México.

Author Contributions: C.H.-G., R.B.-U. and L.G.-V. conceived and designed the experiments; C.H.-G., R.B.-U. performed the experiments; C.H.-G., R.B.-U., B.O.-Q. and L.G.-V. analyzed the data; C.H.-G. and L.G.-V. wrote the paper. All authors have read and approved the final manuscript.

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

References and Notes

1. Zipf, G.K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley Press: Oxford, UK, 1949.
2. Heaps, H.S. *Information Retrieval: Computational and Theoretical Aspects*; Academic Press, Inc.: Cambridge, MA, USA, 1978.
3. Piantadosi, S.T. Zipf's word frequency law in natural language: A critical review and future directions. *Psychon. Bull. Rev.* **2014**, *21*, 1112–1130.
4. Bian, C.; Lin, R.; Zhang, X.; Ma, Q.D.Y.; Ivanov, P.C. Scaling laws and model of words organization in spoken and written language. *EPL* **2016**, *113*, 18002.
5. Shannon, C. Prediction and Entropy of Printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50–64.
6. Shannon, C. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
7. Kalimeri, M.; Constantoudis, V.; Papadimitriou, C.; Karamanos, K.; Diakonos, F.K.; Papageorgiou, H. Word-length entropies and correlations of natural language written texts. *J. Quant. Linguist.* **2015**, *22*, 101–118.
8. Papadimitriou, C.; Karamanos, K.; Diakonos, F.; Constantoudis, V.; Papageorgiou, H. Entropy analysis of natural language written texts. *Phys. A Stat. Mech. Appl.* **2010**, *389*, 3260–3266.
9. Kontoyiannis, I.; Algoet, P.H.; Suhov, Y.M.; Wyner, A.J. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inf. Theory* **1998**, *44*, 1319–1327.
10. Ebeling, W.; Poschel, T. Entropy and Long-Range Correlations in Literary English. *EPL* **1994**, *26*, 241.
11. Chang, M.C.; Yang, A.C.C.; Stanley, H.E.; Peng, C.K. Measuring information-based energy and temperature of literary texts. *Phys. A Stat. Mech. Appl.* **2017**, *468*, 783–789.
12. Rosso, O.A.; Craig, H.; Moscato, P. Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers. *Phys. A Stat. Mech. Appl.* **2009**, *388*, 916–926.
13. Montemurro, M.A.; Zanette, D.H. Universal Entropy of Word Ordering Across Linguistic Families. *PLoS ONE* **2011**, *6*, 1–9.
14. Pincus, S.M. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2297–2301.
15. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **2000**, *278*, H2039–H2049.
16. Ocak, H. Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy. *Expert Syst. Appl.* **2009**, *36*, 2027–2036.
17. Costa, M.; Goldberger, A.L.; Peng, C.K. Multiscale entropy analysis of biological signals. *Phys. Rev. E* **2005**, *71*, 021906.
18. Guzman-Vargas, L.; Ramírez-Rojas, A.; Angulo-Brown, F. Multiscale entropy analysis of electroseismic time series. *Nat. Hazards Earth Syst. Sci.* **2008**, *8*, 855–860.
19. Costa, M.; Peng, C.; Goldberger, A.L.; Hausdorff, J.M. Multiscale entropy analysis of human gait dynamics. *Phys. A Stat. Mech. Appl.* **2003**, *330*, 53–60.
20. Guzmán-Vargas, L.; Ramírez-Rojas, A.; Hernández-Pérez, R.; Angulo-Brown, F. Correlations and variability in electrical signals related to earthquake activity. *Phys. A Stat. Mech. Appl.* **2009**, *388*, 4218–4228.

21. Costa, M.; Goldberger, A.L.; Peng, C.K. Multiscale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.* **2002**, *89*, 068102.
22. Schürmann, T.; Grassberger, P. Entropy estimation of symbol sequences. *Chaos Interdiscip. J. Nonlinear Sci.* **1996**, *6*, 414–427.
23. Jiménez-Montano, M.A.; Ebeling, W.; Pohl, T.; Rapp, P.E. Entropy and complexity of finite sequences as fluctuating quantities. *Biosystems* **2002**, *64*, 23–32.
24. Eckmann, J.P.; Ruelle, D. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.* **1985**, *57*, 617–656.
25. Grassberger, P.; Procaccia, I. Estimation of the Kolmogorov entropy from a chaotic signal. *Phys. Rev. A* **1983**, *28*, 2591–2593.
26. Hamming, R.W. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **1950**, *29*, 147–160.
27. Humeau-Heurtier, A. The Multiscale Entropy Algorithm and Its Variants: A Review. *Entropy* **2015**, *17*, 3110–3123.
28. Evaluating the Irregularity of Natural Languages. Available online: http://figshare.com/projects/Evaluating_the_irregularity_of_natural_languages/25036 (accessed on 27 September 2017).
29. We repeated our procedure for more books (for details see supporting information online at [28]). For different segments with different sizes, we observed that the standard deviation was very small for most of the languages. These particular results and additional discussions will be published elsewhere.
30. Duda, R.; Hart, P. *Pattern Classification and Scene Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 1973.
31. Wu, S.D.; Wu, C.W.; Lin, S.G.; Wang, C.C.; Lee, K.Y. Time Series Analysis Using Composite Multiscale Entropy. *Entropy* **2013**, *15*, 1069–1084.
32. Wu, S.D.; Wu, C.W.; Lin, S.G.; Lee, K.Y.; Peng, C.K. Analysis of complex time series using refined composite multiscale entropy. *Phys. Lett. A* **2014**, *378*, 1369–1374.
33. Ebeling, W.; Neiman, A. Long-range correlations between letters and sentences in texts. *Phys. A Stat. Mech. Appl.* **1995**, *215*, 233–241.
34. Altmann, E.G.; Cristadoro, G.; Esposti, M.D. On the origin of long-range correlations in texts. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 11582–11587.