entropy

MDPI

Article

Recurrence Networks in Natural Languages

Edgar Baeza-Blancas [1,2], Bibiana Obregón-Quintana [3], Candelario Hernández-Gómez [1], Domingo Gómez-Meléndez [4], Daniel Aguilar-Velázquez [2], Larry S. Liebovitch [5,6,7] and Lev Guzmán-Vargas [2,*]

1 Departamento de Física, Escuela Superior de Física y Matemáticas, Ciudad de México 07738, Mexico; blancasbef@gmail.com (E.B.-B.); hernandezgomez2010@gmail.com (C.H.-G.)
2 Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, Instituto Politécnico Nacional, Ciudad de México 07340, Mexico; zafskumo@hotmail.com
3 Facultad de Ciencias, Univesidad Nacional Autónoma de México, Ciudad de México 04510, Mexico; b.obregon.q@gmail.com
4 Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Zacatecas 98000, Mexico; domag5@hotmail.com
5 Department of Physics, Queens College, City University of New York, New York, NY 11367, USA; larry.liebovitch@qc.cuny.edu
6 Advanced Consortium on Cooperation, Conflict, and Complexity (AC4), Earth Institute, Columbia University, New York, NY 10027, USA
7 Graduate Center, City University of New York, New York, NY 10016, USA
* Correspondence: lguzmanv@ipn.mx; Tel.: +52-55-5729600 (ext. 56873)

Received: 7 April 2019; Accepted: 17 May 2019; Published: 23 May 2019

Abstract: We present a study of natural language using the recurrence network method. In our approach, the repetition of patterns of characters is evaluated without considering the word structure in written texts from different natural languages. Our dataset comprises 85 ebookseBooks written in 17 different European languages. The similarity between patterns of length m is determined by the Hamming distance and a value r is considered to define a matching between two patterns, i.e., a repetition is defined if the Hamming distance is equal or less than the given threshold value r. In this way, we calculate the adjacency matrix, where a connection between two nodes exists when a matching occurs. Next, the recurrence network is constructed for the texts and some representative network metrics are calculated. Our results show that average values of network density, clustering, and assortativity are larger than their corresponding shuffled versions, while for metrics like such as closeness, both original and random sequences exhibit similar values. Moreover, our calculations show similar average values for density among languages which that belong to the same linguistic family. In addition, the application of a linear discriminant analysis leads to well-separated clusters of family languages based on based on the network-density properties. Finally, we discuss our results in the context of the general characteristics of written texts.

Keywords: recurrence networks; natural languages; patterns repetition

1. Introduction

During pastIn recent decades many studies have pointing pointed out the complexity and organizational properties of natural languages, specially especially for written texts. These studies have attracted the attention of researchers from different areas of science who use different approaches to tackle the complexity of natural languages [1–6]. From a morphological point of view, languages are classified in analytics and synthetics. The key difference between one and othersthem is the use of the forms of a lexeme to build words and sentences. In this context, many methods developed in the study

of time-series analysis and network science are potentially suitable to evaluate this kind of structural differences [7–15].

Of particular importance are methods based on based on the quantification of recurrence features from systems whose phase phase-space representation may provide information about invariant properties of particular configurations [11]. For instance, the graphical representation of the recurrence plots permits to visually identifyvisual identification of specific signatures of periodicities or cycles, and these signatures can be quantified in terms of fundamental invariant properties of a dynamical system [16]. In this context, Marwan et al. [17] introduced the concept of recurrence network, as an extension of the recurrence plot analysis, where the recurrence matrix is used to obtain the adjacency matrix, which contains the information of the connectivities. The characterization of the recurrence networks provides additional information of to that obtained from the standard recurrence quantification analysis and permits a direct evaluation of the relatedness of vectors defined in an m-dimensional space [18].

The study of language from a dynamical perspective have has been addressed by researchers from different areas of science, ranging from natural natural-language processing to network and information theory [19–22]. For instance, distributions of distances between successive occurrences of a specific word are well described by a stretched exponential, indicating the presence of memory [23], while other studies have considered the recurrence of words within sentences or paragraphs to estimate the correlation dimension of sets of paragraphs (discourse) [24].

We are interested in evaluating the recurrence of patterns (sequences of characters) of a given length which may provide an alternative approach to look into the differences and similarities, between different natural languages. In particular, we resort to methods like such as correlation dimension and recurrence networks for a direct quantitative way of measuring the level of recurrence with respect to the random configuration of the texts [11,25]. Our goal is to analyze the recurrence of patterns along the text, by examining the spatio-temporal organization of these patterns from a network science perspective. Recently, we reported the application of methods like such as the approximate entropy to the study of irregularities displayed by some natural languages [26–28]. In our approach, we consider the similarity between two patterns of length m based on based on the Hamming distance among them. We define a distance (Hamming distance h) and define a "matching" between two patterns if the distance is equal or less than a given value r. In this way, we are able tocan construct the corresponding adjacency matrix, where a connection between two nodes (patterns) exists when a matching occurs. The recurrence network is constructed for texts from different European languages and some representative network metrics are calculated. First, we study the scaling behavior between the density of the network, a measure closely related to the correlation dimension, and the Hamming distance. Our calculations of several network metrics show similarities between languages which belong to the same linguistic family and differences among the linguistic families as well. These similarities and differences explored here have not yet been reported by any other methodology based on based on nonlinear dynamics. We also compare the values of these network metrics of actual data with their corresponding randomized versions. The paper is organized as follows. In Section 2, the method to construct the recurrence networks and the dataset areis described. The results and discussion are presented in Section 3. Finally, some concluding remarks are given in Section 4.

2. Methods

2.1. Recurrence Networks

One of the characteristics frequently observed in dynamical systems is the recurrence of states, identified as states or configurations which become close to previous ones after some time [11,17]. The Hamming distance h is the number of distinct characters when two patterns of the same length m are compared. In our approach, we consider a text with length N and construct $n = N - m + 1$ subseries (patterns) x_i^m of length m. Next, we compare each of these patterns one by one and establish

a matching if the Hamming distance is less than or equal to a tolerance parameter r. More formally, these recurrences are used to construct a recurrence matrix, defined as:

$$R_{ij} = \Theta(r - h(x_i^m, x_j^m)), \tag{1}$$

where $\Theta(-)$ represents the Heaviside function and $h(x_i^m, x_j^m)$ the Hamming distance. The adjacency matrix associated with the recurrence network is given by $A_{ij} = R_{ij} - \delta_{ij}$, with δ_{ij} the Kronecker delta. Thus, each pattern represents a node, and a connection (link) is defined if there is a match. For instance, in Table 1 we show a simple example of the matrix R_{ij} for the Hamlet's soliloquy: *To_be_or_not_to_be*. In this example, the threshold value $r = 2$ is used to illustrate the construction of the recurrence matrix. In general, it is expected that for very regular texts, the number of repetitions subjected to tolerance r would be bigger compared to the case of a very irregular sequence of symbols where a matching is quite difficult unlikely to occur.

Table 1. Recurrence symmetric matrix for the beginning of Hamlet's famous soliloquy: To-be-or-not-to-be. Here $N = 18$ and we set $m = 3$. The resulting matrix has 16 rows and columns.

$r = 2$	**To_**	**o_b**	**_be**	**be_**	**e_o**	**_or**	**or_**	**r_n**	**_no**	**not**	**ot_**	**t_t**	**_to**	**to_**	**ob_**	**_be**
To_	**1**	0	0	**1**	0	**1**	**1**	0	0	**1**	**1**	**1**	0	**1**	0	0
o_b	0	**1**	0	0	**1**	0	**1**	**1**	0	0	**1**	**1**	0	0	**1**	0
_be	0	0	**1**	0	0	**1**	0	0	**1**	0	0	0	**1**	0	0	**1**
be_	**1**	0	0	**1**	0	0	**1**	0	0	0	**1**	0	0	**1**	0	0
e_o	0	**1**	0	0	**1**	0	0	**1**	**1**	0	0	**1**	**1**	0	**1**	0
_or	0	0	**1**	0	0	**1**	0	0	**1**	**1**	0	0	**1**	**1**	0	**1**
or	**1**	**1**	0	**1**	0	0	**1**	0	0	0	**1**	0	0	**1**	**1**	0
r_n	0	**1**	0	0	**1**	0	0	**1**	0	0	0	**1**	0	0	**1**	0
_no	0	0	**1**	0	**1**	**1**	0	0	**1**	0	0	0	**1**	0	0	**1**
not	**1**	0	0	0	0	**1**	0	0	0	**1**	0	**1**	0	**1**	0	0
ot_	**1**	**1**	0	**1**	0	0	**1**	0	0	0	**1**	0	**1**	**1**	**1**	0
t_t	**1**	**1**	0	0	**1**	0	0	**1**	0	0	0	**1**	0	**1**	**1**	0
_to	0	0	**1**	0	**1**	**1**	0	0	**1**	0	0	0	**1**	0	0	**1**
to_	**1**	0	0	**1**	0	**1**	**1**	0	0	**1**	**1**	**1**	0	**1**	0	0
o_b	0	**1**	0	0	**1**	0	**1**	**1**	0	0	**1**	**1**	0	0	**1**	0
_be	0	0	**1**	0	0	**1**	0	0	**1**	0	0	0	**1**	0	0	**1**

2.2. Network Metrics

During past yearsIn recent years, many studies focused on complex systems have used the network approach to characterize spatial and temporal organization of systems composed of interacting units [29]. Network measures may be useful to understand diverse properties of large sequences of symbols, which are transformed to matrices like like in the case of recurrence analysis [17,18]. Here, we listed some basic network metrics:

- Density (ρ): The density of a network is defined as:

$$\rho = \frac{2g}{n(n-1)}, \tag{2}$$

with g the number of actual connections and n is the number of nodes (patterns). A value of ρ close to 1 denotes an almost complete graph and ρ close to 0 indicates a poorly connected network.

- Closeness centrality (K_c): Measures the centrality of a given node in the network, defined as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph [29],

$$K_c = \frac{n}{\Sigma_j d_{ij}}, \tag{3}$$

where d_{ij} denotes the distance from node i to node j.

- Clustering coefficient (C_i): Measures the degree of transitivity in connectivity amongstamong the nearest neighbors of a node i [29]. In recurrence terms, C_i represents the extent to which neighbors of a node (pattern) i are also recurrent amongstamong themselves. Specifically, C_i is given by,

$$C_i = \frac{2E_i}{k_i(k_i - 1)}, \tag{4}$$

where E_i is the number of links between the k_i neighbors of the node i.

- Average nearest-neighbor degree ($\bar{k}_{nn,i}$): This measure allows us to see the mean preference in connectivity of a given node [30–32]. The behavior of this quantity as a function of the node's degree, reveals whether high-degree nodes connect with other equally high-degree ones (assortativity), or high-degree nodes preferentially connect to low-degree ones (dissortativity) [29]. For unweighted networks, $\bar{k}_{nn,i}$ is calculated as:

$$\bar{k}_{nn,i} = \frac{1}{k_i} \sum_{j=1}^{N} A_{ij} k_j, \tag{5}$$

where k_i is the node's degree, A_{ij} represents the adjacency matrix and N is the number of nodes.

- Assortative mixing coefficient by degree (A_r): This measure quantifies the tendency observed in networks that nodes with many connections are connected to other nodes with many (or a few) connections [33]. Formally, the coefficient is given by,

$$A_r = \frac{\sum_{ij} A_{ij}(k_i - \mu)(k_j - \mu)}{\sum_{ij} A_{ij}(k_i - \mu)^2}, \tag{6}$$

where $\mu = \frac{\sum_{ij} A_{ij} k_i}{\sum_{ij} A_{ij}}$. For perfectly assortative networks, the coefficient reaches a maximum value of 1, whereas a minimum value of -1 is observed for perfectly dissortative ones.

3. Results

We constructed recurrence networks from texts described in https://figshare.com/articles/Recurrence_networks_in_natural_languages/7885376 [34]. The corpus is comprised of 85 books from 17 different languages, corresponding to 4 linguistic families (Germanic, Romance, Slavic, and Uralic). In order toTo validate our method for relatively short sequences, in our calculations we restrict ourselves to segments with 15,000 symbols and repeat the calculations for 5 segments of this length. In our case we have kept the punctuation marks and the space mark as symbols.

Prior to the presentation of network metrics results, we explored the behavior of the recurrence-network connectivity in terms of the tolerance parameter r for different values of the pattern length m. The recurrence rate [35] is a useful measure to quantify the connectivities and it is related to the correlation sum defined in the context of correlation dimension [25] and to the density in the context of networks. We notice that given the length of the alphabet L in one natural language, the number of manners that r-number of discrepancies that may occur is given by $r_0 = L^r$, where r_0 represents the total number of permutations with repetitions. Figure 1 shows the behavior between the density ρ and r in a log-linear plane for a text in the English language. A linear behavior is observed, where the slope is given by the exponent value $d = 1.3 \pm 0.02$ for $m = 10$. Next, we calculated several network metrics for the texts from different languages. As we stated above, in our calculations we considered 5 segments

with $N = 15{,}000$ elements and we set $m = 5$, which is a value that roughly corresponds to the mean word length in several languages [26,36–38]. The threshold error value is set to $r = 2$. The results for the density ρ are presented in Figure 2a, where languages were grouped according to the linguistic family to which they belong. We observe that the Germanic family exhibits high values of density, followed by the Romance family, whereas the Slavic and the Uralic ones are represented by lower values. Here, a high value of density indicates that the number of recurrences is higher than recurrences in other families. For each language, we also generate a surrogate text sequence by shuffling the characters of the original text and show its corresponding value of the density. These results clearly indicate that the Germanic family is the one that is more separated from their corresponding random cases, while both the Slavic and the Uralic are located close to the random configuration.

The results of local structure represented by the clustering coefficient are depicted in Figure 2b. In this case, all the languages exhibit a similar intermediate average value for this measure, suggesting that this value of local structure is likely "universal" across different languages, i.e., the probability that neighbors of a pattern are also neighbors with each other is somehow intermediate. We noticed that for shuffled texts, the values of the average clustering coefficient are smaller (around 0.2) than the values of the original ones. The identification of this intermediate value of the clustering may help to understand the balanced local and global structures.

Moreover, the average closeness values are quite similar across different linguistic families, except data from the Slavic family and Hungarian, which display values below 0.27, indicating that their average farness (inverse distance) between nodes are smaller than values from the other languages, i.e., large distances from a node (pattern) to all other nodes are more likely to observe in recurrence networks from Slavic and Hungarian texts (see Figure 2c). Unlike the clear difference observed between the two previous network metrics and their corresponding shuffled version, the average closeness values obtained from randomized data mostly overlap with original calculations of closeness, confirming that values of shortest paths between nodes almost do not change after randomizing texts. This result points out that both original and random recurrence-pattern networks share one of the recognized small-world properties [39], i.e., while distance between nodes tends to be small in both networks, the clustering (local structure) is higher for original texts (see panels (b) and (c) in Figure 2).

The mixing pattern by degree represents an alternative to capture the tendency of correlations in terms of connectivities. For instance, assortative mixing indicates that there are positive correlations between node's degree, i.e., nodes with many connections tend to connect to other nodes that also have many connections, while negative correlations (dissortative mixing) indicate that nodes with many connections tend to connect to other nodes that have few few connections. If no correlation by degree is identified, it is said that there is no mixing pattern. The results of mixing pattern are shown in Figure 2d. We found an assortative mixing pattern for all the languages under study with values between 0.5 to 0.7, which are relatively higher than typical values reported for other systems [33,40].

In the same direction, we also evaluated the behavior of the average nearest- neighbor connectivity as a function of the degree.

Figure 3 shows the results of the behavior of k_{nn} vs. k for original and randomized texts. We observe a scaling behavior for all languages of the form $k_{nn} \sim k^{\delta}$, with $\delta \approx 0.49 \pm 0.03$ for original texts, while for random ones, the exponent is $\delta \approx 0.47 \pm 0.02$. It is worth to It is worth remarking that for low low-degree values, the values of k_{nn} from original and random texts are markedly different from each other.

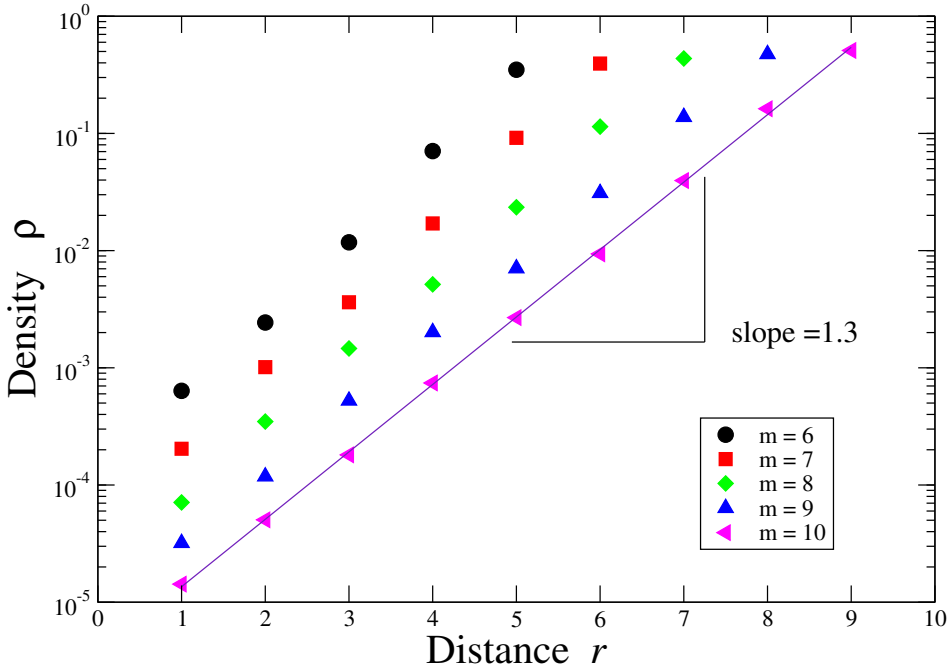

Figure 1. Log-linear plot of density ρ vs. the distance r for several values of the pattern length m. Here we show the cases $m = 6, 7, 8, 9, 10$ and r runs from 1 to r_{max}, where $r_{max} = m - 1$. The fit corresponds to the case $m = 10$, which yields to $d \approx 1.3$.

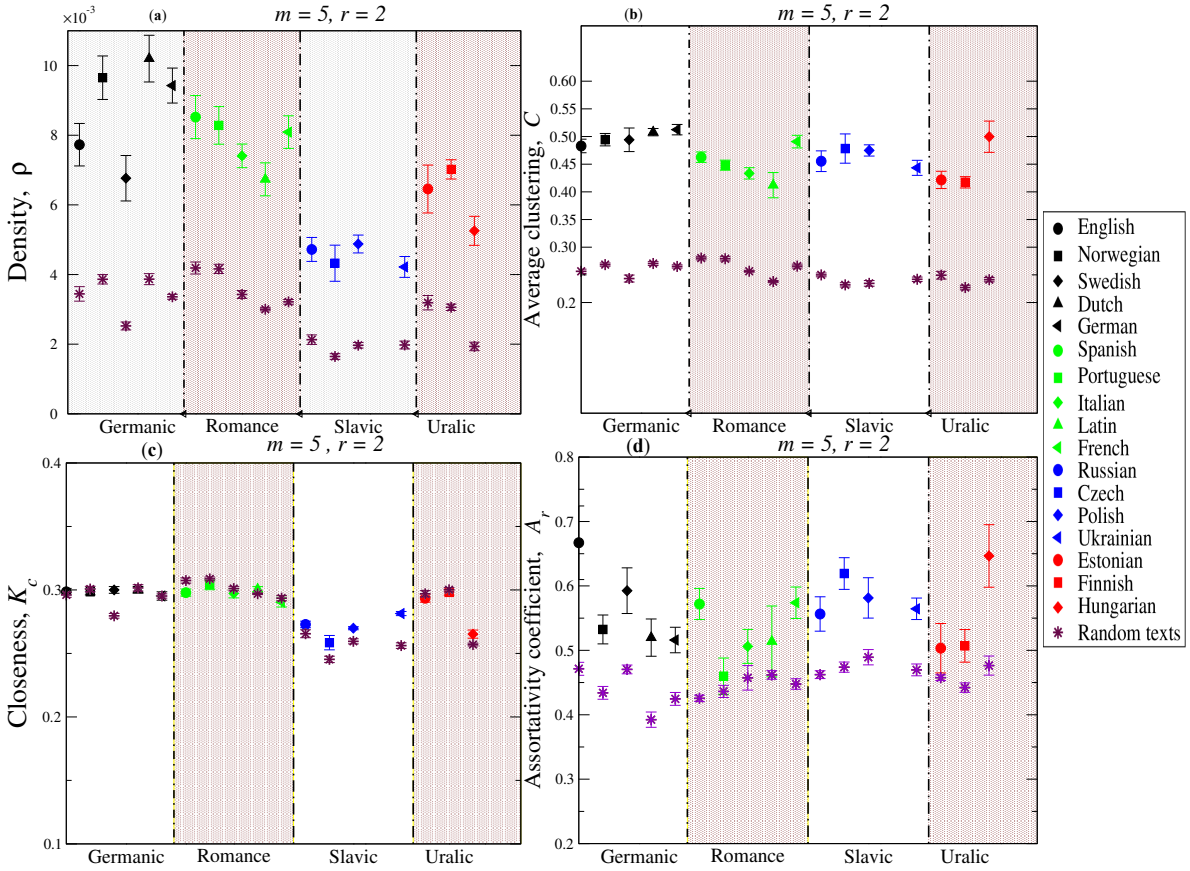

Figure 2. Representative metrics of recurrence-pattern networks for different languages. (**a**) Density for languages grouped by linguistic families. (**b**) Average clustering coefficient $\mathcal{C}$. (**c**) Closeness centrality. (**d**) Assortativity coefficient. Vertical bars indicate the standard deviation of the data.

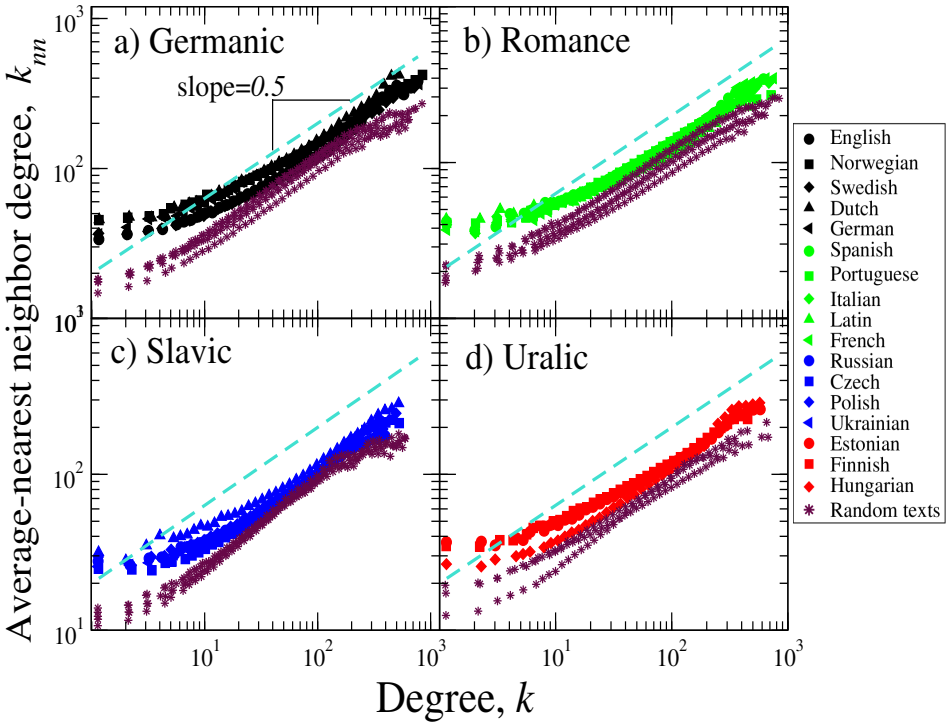

Figure 3. Mean nearest-neighbor connectivity as a function of the degree for (**a**) Germanic, (**b**) Romance, (**c**) Slavic, and (**d**) Uralic linguistic families. For each language, we also show the values of k_{nn} corresponding to shuffled texts. A scaling behavior is observed for all cases of the form $k_{nn} \sim k^{\delta}$. We estimate the scaling exponent for degree values $10 < k < 500$, yielding the average values $\bar{\delta} \approx 0.49$ and $\bar{\delta_r} \approx 0.47$ for the original and random data, respectively. As a guide for the eye, the dashed line corresponds to the slope = 0.5.

Finally, in order toto provide a direct comparison between different languages based on based on network-metric values, we applied the Fisher's linear discriminant analysis (LDA) [41] to the density values reported in Figure 2. This technique is very useful to determine if the density could potentially classify languages into the linguistic families they belong to. For this analysis we considered the average density values for each language. Then, the data were projected down to a two-dimensional scatter plot presented in Figure 4a. We observe a separation between clusters formed by languages that belong to the same linguistic family, except the case of the Uralic and Romance families, which are divided into two clusters each one. For a better evaluation of the separation of the clusters provided by the discriminant analysis, we also applied the *k*-nearest- neighbor classification method using as input the results provided by the LDA method. The results are presented in Figure 4b,c. The performance of the system reports that the classifier correctly guessed 89.4% of the times by using $m = 5, 6, 7$ as input information.

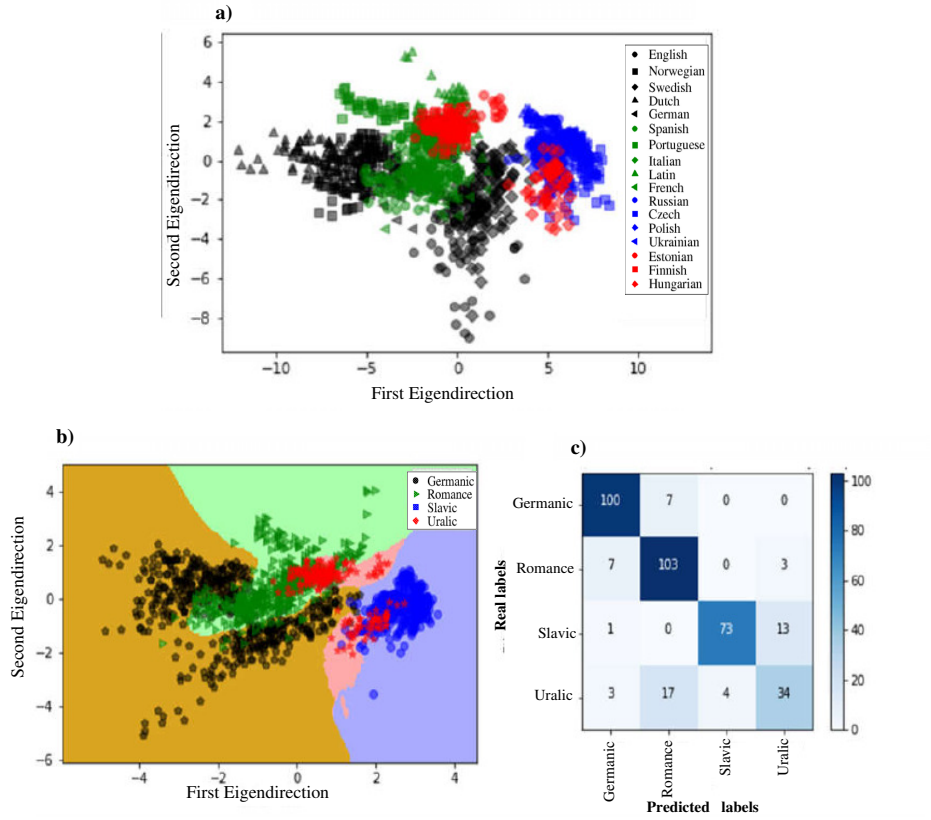

Figure 4. Results of classification analysis applied to European languages. (**a**) Results of the linear discriminant method. Here we show the projection of density values from pattern lengths $m = 5, 6, 7$. For each m-value and for each language, we considered ten segments with length 10^4 to obtain ten ρ values. Next, languages were labeled in classes according to the linguistic family to which they belong (Romance, Germanic, Slavic, Uralic). (**b**) Results of the application of the k-nearest- neighbor classification method to data in panel a) but assigning the same label to languages of the same family. We used $k = 20$ neighbors in the classifier. We observe that the families are segregated, except in the case of the Uralic family, which led to two disjoint regions. (**c**) Results of the confusion matrix. The system makes a clear distinction between almost all family languages, except the case of Uralic, where we observe a problem distinguishing this family from Slavic and Romance.

4. Discussion and Conclusions

Our results point out that network metrics applied to recurrence dynamics permit to characterizethe characterization of the natural language at different levels. At the local level, metrics like such as clustering coefficient revealed that the local structure is quite similar across different languages, suggesting a general property of organization in natural languages. At the global level, information likesuch as mixing patterns and closeness also indicate that there some similar features, but other metrics like such as the density assign different values to different linguistic families while languages which belong to the same family exhibit similar values. Notably, the profiles of correlations in connectivities of nearest neighbors appear in a similar way across different linguistic families. The results we report here are in general concordance with previous studies focused on phonological networks [42–44], which reported positively correlated behavior for assortative mixing and relatively low value for clustering coefficients. In our case, we observed higher values for both mixing coefficient and clustering coefficients compared to the values reported in phonological networks [13,42]. The small-world structure was also observed in the recurrence-language networks. This result is consistent with previous studies which argue that lexical retrieval processes are more "efficient" in the sense that a rapid and robust search is optimal for some network configurations [42,44].

Although our approach is based on based on repetition of patterns without considering specific elements like such as words, and the languages come from a diverse range of linguistic families, the properties of recurrence networks suggest that local and global structures display similarities and differences between languages, opening the possibility of a quantitative evaluation between them. Furthermore, real texts exhibited important differences in the network structure compared to texts obtained from randomizations, i.e., natural languages have more complex structure compared to shuffled sequences. It is also noticeable that for some network-based metrics like such as density, some differences are clearly identified between linguistic familiesfamilies, but additional analyses are needed in this direction. The application of the linear discriminant analysisLDA together with the classification method to the network-based values, revealed that some of the languages are segregated but others are not distinguishable from each other. In summary, our recurrence- network procedure have has revealed additional organizational properties of language, which confirms that there exist similarities in network properties, and some differences emerge between languages that belong to different linguistic families. We remark that these network analysisthese network analyses have not been previously reported for natural languages. Moreover, our results reinforcesreinforce the idea that many aspects of language can be evaluated from a network perspective. Finally, we point out that additional studies are needed to fully characterize the recurrence-network properties of natural language.

Author Contributions: E.B.-B., B.O.-Q., L.S.L. and L.G.-V. analyzed the data; E.B.-B., B.O.-Q., C.H.-G., D.G.-M., D.A.-V., and L.S.L. contributed analysis tools; L.S.L. and L.G.-V. and contributed to the conception and design of the study and data interpretation, L.S.L., and L.G.-V. wrote the paper.

Funding: This work was partially supported by COFAA-IPN, EDI-IPN, Conacyt-México and PAPIIT-UNAM (IA303418). L.G.-V. carried out this work during a leave supported by COTEBAL-IPN.

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

References

1. Zipf, G.K. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*; Houghton Mifflin: Boston, MA, USA, 1935.
2. Grzybek, P. History and methodology of word length studies. In *Contributions to the Science of Text and Language. Text, Speech and Language Technology*; Grzybek, P., Ed.; Springer: Dordrecht, The Netherlands, 2006; Volume 31, pp. 15–90.
3. Piantadosi, S.T.; Tily, H.; Gibson, E. Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 3526–3529. [CrossRef]
4. Solé, R.V.; Corominas-Murtra, B.; Valverde, S.; Steels, L. Language networks: Their structure, function, and evolution. *Complexity* **2010**, *15*, 20–26. [CrossRef]
5. Seoane, L.F.; Solé, R. The morphospace of language networks. *Sci. Rep.* **2018**, *8*, 10465. [CrossRef]
6. Rêgo, H.H.A.; Braunstein, L.A.; D'Agostino, G.; Stanley, H.E.; Miyazima, S. When a Text Is Translated Does the Complexity of Its Vocabulary Change? Translations and Target Readerships. *PLoS ONE* **2014**, *9*, e110213. [CrossRef]
7. Kosmidis, K.; Kalampokis, A.; Argyrakis, P. Language time series analysis. *Physica A* **2006**, *370*, 808–816. [CrossRef]
8. Lacasa, L.; Luque, B.; Ballesteros, F.; Luque, J.; Nuño, J.C. From time series to complex networks: The visibility graph. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 4972–4975. [CrossRef]
9. Luque, B.; Lacasa, L.; Ballesteros, F.; Luque, J. Horizontal visibility graphs: Exact results for random time series. *Phys. Rev. E* **2009**, *80*, 046103. [CrossRef] [PubMed]
10. Ausloos, M. Generalized Hurst exponent and multifractal function of original and translated texts mapped into frequency and length time series. *Phys. Rev. E* **2012**, *86*, 031108. [CrossRef]
11. Donner, R.V.; Zou, Y.; Donges, J.F.; Marwan, N.; Kurths, J. Recurrence networks—a novel paradigm for nonlinear time series analysis. *New J. Phys.* **2010**, *12*, 033025. [CrossRef]
12. Rodriguez, E.; Aguilar-Cornejo, M.; Femat, R.; Alvarez-Ramirez, J. Scale and time dependence of serial correlations in word-length time series of written texts. *Physica A* **2014**, *414*, 378 – 386. [CrossRef]

13. Arbesman, S.; Strogatz, S.H.; Vitevitch, M.S. Comparative Analysis of Networks of Phonologically Similar Words in English and Spanish. *Entropy* **2010**, *12*, 327. [CrossRef]
14. De Arruda, H.F.; Marinho, V.Q.; Costa, L.d.F.; Amancio, D.R. Paragraph-based representation of texts: A complex networks approach. *Inf. Process. Manag.* **2019**, *56*, 479–494. [CrossRef]
15. Susuki, S.; Hirata, Y.; Aihara, K. Definition of distance for marked point process data and its application to recurrence plot-based analysis of exchange tick data of foreign currencies. *Int. J. Bifurcat. Chaos* **2010**, *20*, 3699–3708. [CrossRef]
16. Trulla, L.; Giuliani, A.; Zbilut, J.; Webber, C., Jr. Recurrence quantification analysis of the logistic equation with transients. *Phys. Lett. A* **1996**, *223*, 255–260. [CrossRef]
17. Marwan, N.; Donges, J.F.; Zou, Y.; Donner, R.V.; Kurths, J. Complex network approach for recurrence analysis of time series. *Phys. Lett. A* **2009**, *373*, 4246–4254. [CrossRef]
18. Zou, Y.; Donner, R.V.; Marwan, N.; Donges, J.F.; Kurths, J. Complex network approaches to nonlinear time series analysis. *Phys. Rep.* **2019**, *787*, 1–97. [CrossRef]
19. Liu, H.; Cong, J. Language clustering with word co-occurrence networks based on parallel texts. *Sci. Bull.* **2013**, *58*, 1139–1144. [CrossRef]
20. Abramov, O.; Mehler, A. Automatic Language Classification by means of Syntactic Dependency Networks. *J. Quant. Linguist.* **2011**, *18*, 291–336, [CrossRef]
21. Martinčić-Ipšić, S.; Margan, D.; Meštrović, A. Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. *Physica A* **2016**, *457*, 117–128. [CrossRef]
22. Montemurro, M.A.; Zanette, D.H. Universal Entropy of Word Ordering across Linguistic Families. *PLoS ONE* **2011**, *6*, e19875. [CrossRef]
23. Altmann, E.G.; Pierrehumbert, J.B.; Motter, A.E. Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words. *PLoS ONE* **2009**, *4*, e7678. [CrossRef]
24. Doxas, I.; Dennis, S.; Oliver, W.L. The dimensionality of discourse. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 4866–4871. [CrossRef]
25. Grassberger, P. Generalized dimensions of strange attractors. *Phys. Lett. A* **1983**, *97*, 227–230. [CrossRef]
26. Hernández-Gómez, C.; Basurto-Flores, R.; Obregón-Quintana, B.; Guzmán-Vargas, L. Evaluating the Irregularity of Natural Languages. *Entropy* **2017**, *19*, 521. [CrossRef]
27. Pincus, S.M. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 2297–2301. [CrossRef]
28. 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. [CrossRef]
29. Newman, M.E.J. *Networks: An Introduction*; Oxford University Press: Oxford, UK, 2010.
30. Pastor-Satorras, R.; Vázquez, A.; Vespignani, A. Dynamical and Correlation Properties of the Internet. *Phys. Rev. Lett.* **2001**, *87*, 258701. [CrossRef]
31. Maslov, S.; Sneppen, K. Specificity and Stability in Topology of Protein Networks. *Science* **2002**, *296*, 910–913. [CrossRef] [PubMed]
32. Barrat, A.; Barthélemy, M.; Pastor-Satorras, R.; Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 3747–3752. [CrossRef]
33. Newman, M.E.J. Mixing patterns in networks. *Phys. Rev. E* **2003**, *67*, 026126. [CrossRef]
34. Available online: https://figshare.com/articles/Recurrence_networks_in_natural_languages/7885376 (accessed on 23 May 2019).
35. Marwan, N.; Wessel, N.; Meyerfeldt, U.; Schirdewan, A.; Kurths, J. Recurrence-plot-based measures of complexity and their application to heart-rate-variability data. *Phys. Rev. E* **2002**, *66*, 026702. [CrossRef]
36. Kalimeri, M.; Constantoudis, V.; Papadimitriou, C.; Karamanos, K.; Diakonos, F.K.; Papageorgiou, H. Entropy analysis of word-length series of natural language texts: Effects of text language and genre. *Int. J. Bifurcat. Chaos* **2012**, *22*, 1250223. [CrossRef]
37. 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. [CrossRef]
38. Guzmán-Vargas, L.; Obregón-Quintana, B.; Aguilar-Velázquez, D.; Hernández-Pérez, R.; Liebovitch, L. Word-length correlations and memory in large texts: A visibility network analysis. *Entropy* **2015**, *17*, 7798–7810. [CrossRef]

39. Watts, D.J.; Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **1998**, *393*, 440–442. [CrossRef]
40. Newman, M.E.J. Assortative Mixing in Networks. *Phys. Rev. Lett.* **2002**, *89*, 208701. [CrossRef]
41. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
42. Vitevitch, M.S. What Can Graph Theory Tell Us About Word Learning and Lexical Retrieval? *J. Speech Lang. Hear. Res.* **2008**, *51*, 408–422. [CrossRef]
43. Arbesman, S.; Strogatz, S.H.; Vitevitch, M.S. The structure of phonological networks across multiple languages. *Int. J. Bifurcat. Chaos* **2010**, *20*, 679–685. [CrossRef]
44. Chan, K.Y.; Vitevitch, M.S. Network Structure Influences Speech Production. *Cogn. Sci.* **2010**, *34*, 685–697. [CrossRef]