IJCSS International Journal of Complex Systems in Science

Int. J. Complex Systems in Science
vol. **1(2)** (2011), pp. 221–225

Topological Complexity in Natural and Formal Languages

J.P. Cárdenas[1,2,†], J.C. Losada[1], A. Moreira[2], I.G. Torre[1] and R.M. Benito[1]

[1] *Universidad Politécnica de Madrid, Grupo de Sistemas Complejos, E.T.S.I. Agrónomos. 28040 Madrid, Spain*

[2] *Instituto de Sistemas Complejos de Valparaíso, Subida Artillería 470 Valparaíso, Chile.*

Abstract.

Since the beginning of this century many complex systems have been studied by Complex Network Theory. Social, biological and technological systems have been viewed as networks were nodes represent system elements and edges represent the relationship between them. In this work we study languages from this perspective. Transforming texts in networks of words (i.e. finite strings of letters, symbols, or tokens) we observe that languages display common topological properties with other complex systems. Scaling in the distribution of word connectivity, properties of *small-world* networks, multifractal behaviour, among others properties, have been observed in networks extracted from texts written in Natural and Formal languages. We observe that many of these properties seem to depend on the frequency of words in the text, however other ones seem to be strictly determined by the grammar of language.

Keywords: Complex Networks, Word Networks, Grammar, Language
MSC 2000: 05C82

† **Corresponding author:** jpcardenas.agronomos@upm.es
Received: October 19, 2011
Published: October 24, 2011

1. Introduction

A very active interdisciplinary area of research over the last decade has been the study of complex networks, a generic name which refers to several classes of large graphs with characteristic properties [1, 2]. In this work we are interested in networks of word, from which several kinds have appeared in the literature [3, 4, 5]. The type of network we are concerned with is built by using a particular text: the nodes of the network are the words that appear in it, and they are connected if they are found next to each other (co-occurrence) in some part of the text. Thus, $G(N, E)$ corresponds to the graph of co-occurrence where N is the set of different words of the text and E the set of edges (bidirectional) between them. Now, for G is defined the respective adjacency matrix $A(G)$ for which, if $A_{ij} = 1$, there is an edge between the words i and j, while if $A_{ij} = 0$ the edge does not exist. If the words i and j appear together c times in the text then $A_{ij} = c$. Figure 1 shows a word network obtained from a text conformed by two phrases.

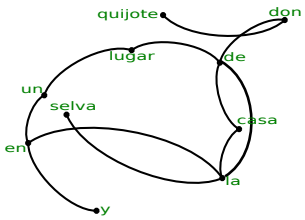

Figure 1: Word network, $G(10, 11)$, corresponding to the text: "*en un lugar de la selva. y en la casa de don quijote*". Repeated words, *en*, *la* and *de*, are more connected than the others. Notice that all the words are written with lower-case letters and we consider (any) punctuation as link cutter.

2. Results

Word networks display properties observed in many complex networks. Left plot of Fig. 2 shows the distribution of connectivities for the words in the network corresponding to the book "Don Quijote de la Mancha". As can be seen, the network presents scaling in the degree distribution, one of the signatures of complex networks [2]. Besides, this graph, as well as other studied, displays properties of *small-world* networks where a high clustering coefficient ($\langle C \rangle \approx 0.20$) and a short path length ($\langle l \rangle \approx 3.5$) appear together. It must be emphasized that the same properties have been observed in networks constructed with the same words and frequency, but randomized (i.e., words

located randomly in different places in the text). This result suggests that some of the observed properties may be dependent on the frequency of words in the text. However, other properties such as the distribution of clustering by degree, $C(k)$, seems to be dependent on the grammar. Right plot of Fig. 2 shows how this distribution is altered in randomized (non grammar) texts.

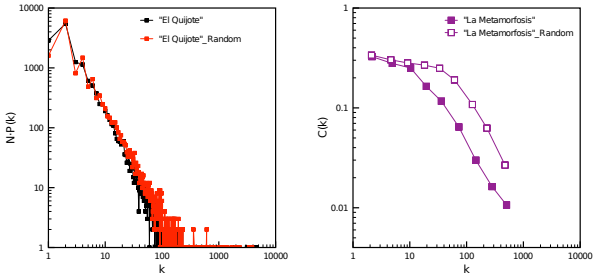

Figure 2: *Left*: Distribution of the number of words with k connections, $N \cdot P(k)$, for "*Don Quijote de la Mancha*" ($N = 3895$). Original text in black and randomized version in red. *Right*: $C(k)$ distribution for "La Metamorfosis". Original text in purple and randomized version in white.

We also analysed the dynamics on the word networks identifying the moment of the first appearance (FTA) of the words in the text. Left plot in the first row of Fig. 3 shows this for "The Universal Declaration of Human Rights" written in different languages. A clear slope define the incorporation of new words in all the languages, however the slope is different for languages with different origin but similar in those with the same origin.

Using the same method to construct networks described in the Introduction we build networks of texts written in Formal languages (e.g., computer programs) where nodes are *tokens*. As in the case of Natural languages, topological properties that depend on tokens frequency appear again. However, a clear difference between Natural and Formal languages is observed in the dynamics of FTA. Figure 3, right plot in the first row, shows that Formal languages are more structured than Natural languages. Besides, randomized Formal languages are completely different to the original versions, but surprisingly similar to Natural languages.

Finally, in order to characterize more deeply the complexity of the networks we apply the "multifractal formalism" [6] with the purpose to search for self-similar structures in the adjacency matrix $A(G)$. Using the box counting method we obtain the *generalized dimensions*, D_q, of the matrix. These dimensions are deeply connected with the thermodynamic formalism of equilibrium

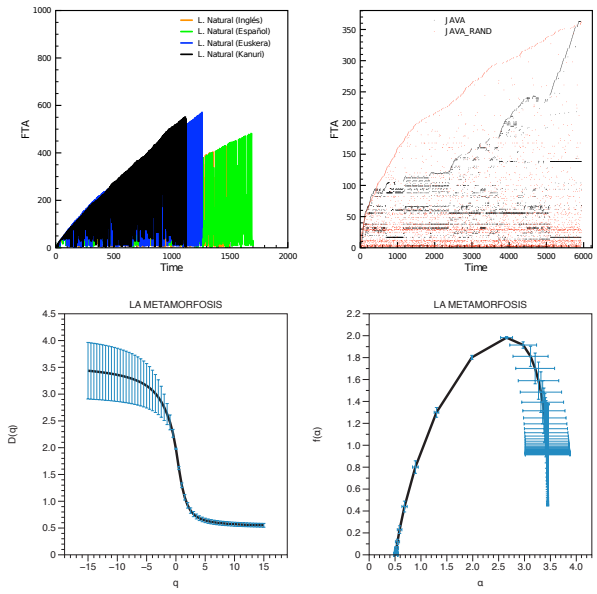

Figure 3: *First row*: Dynamics of FTA in "The Universal Declaration of Human Rights" written in different languages (left) and in a computer program written in JAVA (right). Original program in black and randomized version in red. *Second row*: Generalized dimensions D_q (left) and multifractal spectrum $f(\alpha)$ (right) for box counting method applied to $A(G)$. Standard errors bars in blue.

statistical mechanics through the equation, $D_q = \frac{1}{q-1} \lim_{L\to\infty} \frac{log \sum_i P_i^q(L)}{log L}$, where D_q correspond to scaling for the qth moments of the measure and L the size of the box. The left plot in the second row of Fig. 3 shows D_q for the network of "La Metamorfosis". If D_q were constant this would be a monofractal behaviour, however, as can be seen, the matrix displays multifractal behaviour: the dimensions depend on the q exponent generating a strong change of slope at $q = 0$. The plot of the right side shows the multifractal spectrum $f(\alpha)$ for the same text. According to the formalism, $f(\alpha)$ provides a precise and intuitive description of the multifractal measure in terms of interwoven sets, with singularity strength α, whose Hausdorff dimension is $f(\alpha)$. The convexity of the curve reveals the multifractal nature of $A(G)$. The asymmetry of this curve respect to the maximum, in which for smaller values of α the value of $f(\alpha)$ is 0, denotes the contributions of high-weighted links on the network complexity. Moreover these values show lower errors, scaling better to make the calculation of the Hausdorff dimension.

3. Conclusions

Word networks display complexity, similar to the one observed in other complex systems. Multifractal behaviour and emergent topological properties have been observed in different languages. Some of these properties seem to depend on the word frequency; the presence of similar properties in randomized versions of these texts is an example of this. However, other properties, such as the distribution $C(k)$, seem to depend on the connectivity of words.

The study of the dynamics of word networks shows differences between Natural languages with different origin. Besides, denotes a clear difference between Natural and Formal languages. However, it is necessary to mention that networks of Formal languages show similar topological properties to those observed in networks of Natural languages.

References

[1] M. E. J. Newman, SIAM Review **2**, 167-256 (2003).

[2] M. E. J. Newman, A-L. Barabási and D. Watts, The Structure and Dynamics of Networks Princeton University Press, (2006).

[3] R. Ferrer i Cancho and R.V. Solé, Proceedings of the Royal Societyof London B **286**, 2261-2266 (2001).

[4] Y. Huang, J. Tan and L. Zhang, Journal of Software **4**, 3-10 (2009).

[5] L. Sheng and C. Li, Physica A **388**, 2561-2570 (2009).

[6] A. Chhabra and R.V. Jensen, Physical Review Letters **62**, 1327-1330 (1989).