Network Differences between Normal and Shuffled Texts: Case of Croatian

Domagoj Margan, Sanda Martinčić-Ipšić, and Ana Meštrović

Abstract. This paper is an initial attempt to study the properties of the Croatian word order via complex networks. We present network properties of normal and shuffled Croatian texts for different co-occurrence window sizes and different linkage boundaries. The results of network analysis show that the text shuffling causes the decrease of the network diameter, due to the establishment of previously non-existing links. This indicates that the syntax does play a significant role in the Croatian language, although it is a mostly free word-order language.

Keywords: complex networks, linguistic co-occurrence networks, Croatian corpus, shuffled text, randomized text.

1 Introduction

The complex networks sub-discipline tasked with the analysis of language has been recently associated with the term of linguistic's network analysis. The linguistic network can be based on various language constraints: structure, semantics, syntax dependencies, etc. It has been shown that language networks share various non-trivial topological properties and may be characterized as small-world networks and scale-free networks which are well-known and studied classes of complex networks. Small-world networks [14] have a small average shortest path length and a large clustering coefficient; scale-free networks [4] have power law degree distribution.

Domagoj Margan · Sanda Martinčić-Ipšić · Ana Meštrović
Department of Informatics,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
e-mail: {dmargan,smarti,amestrovic}@uniri.hr

P. Contucci et al. (eds.), *Complex Networks V*,
Studies in Computational Intelligence 549,
DOI: 10.1007/978-3-319-05401-8_26, © Springer International Publishing Switzerland 2014

In the linguistic co-occurrence complex networks properties are derived from the word order in texts. The open question is how the word order itself is reflected in topological properties of the linguistic network. One approach to address this question is to compare networks constructed from normal texts with the networks from randomized or shuffled texts. Since the majority of linguistic network studies have been performed for English, it is important to check whether the same properties hold for Croatian language as well. In this context the study of the Croatian language is notably behind other European languages [1]. So far, there have been only sporadic efforts to model the phenomena of the Croatian language through complex networks. Croatian is a highly flective Slavic language and words can have 7 different cases for singular and 7 for plural, genders and numbers. The Croatian word order is mostly free, especially in non-formal writing. These features are positioning Croatian among morphologicaly rich and free word-order languages.

In this paper we address the problem of Croatian text complexity by constructing the linguistic co-occurrence networks form two types of corpora: a) Croatian original texts, b) Croatian word-level shuffled texts. For the construction of the networks we varied two different criteria: a) the co-occurrence window size, b) the delimiters for limiting the linkage of the words only to the borders of a sentence.

Section 2 presents an overview of related work on complex network analysis of randomized texts. In Section 3 we define measures for the network structure analysis. In Section 4 we present the construction of eight different co-occurrence networks. The network measurements are in Section 5. In the final Section, we elaborate the obtained results and make conclusions regarding future work.

2 Related Work

Some of the early work related to the analysis of random texts dates to 1992, when Li [8] showed that the distribution of word frequencies for randomly generated texts is very similar to Zipf's law observed in natural languages such as in English. Thus, the feature of being a scale-free network does not depend on the syntactic structure of the language. Watts and Strogatz [14] showed that the network formed by the same amount of nodes and links but only establishing links by choosing pairs of nodes at random has a similar small network distance measures as in the original one. Caldeira *et al.* [5] analyzed the role played by the word frequency and sentence length distributions to the undirected co-occurrence network structure based on shuffling. Shuffling procedures were conducted either on the texts or on the links. Liu and Hu [9] discussed whether syntax plays a role in the complexity measures of a linguistic network. They built up two random linguistic networks based on syntax dependencies and compared the complexity of non-syntactic and syntactic language networks. Masucci and Rodgers showed [11, 12] that the power law distribution holds when they randomized the words in the text. Thus, they showed that degree distribution is not the best measure of the self-organizing nature of weighted linguistic networks. Due to the equivalence between frequency and strength of a node, shuffled texts obtain the same degree distribution, but lose all the syntactic structure. They have analyzed the differences between the statistical properties of a real and a shuffled weighted network and showed that the scale-free degree distribution and the scale-free weight distribution are induced by the scale-free strength distribution. They defined a measure, the node selectivity, that can distinguish a real network from a shuffled network. Krishna *et al.* [7] studied the effect of linguistic constraints on the large scale organization of language. They described the properties of linguistic co-occurrence networks with the randomized words. These properties were compared to those obtained for a network built over the original text. It is observed that the networks from randomized texts also exhibit small-world and scale-free characteristics.

Preliminary results on Croatian co-occurrence networks presented in [10] point out that the increase of the co-occurrence window size is followed by a decrease in diameter, average path shortening and, expectedly, the condensing of the average clustering coefficient. The stopwords removal causes the same effect. When comparing Croatian literature networks to networks from other languages such as English and Italian [3] some expected universalities such as small-world properties are shown, but there are still some differences. The Croatian language exhibits a higher path length than English and Italian language which can be caused by the mostly free word order nature of Croatian.

3 The Network Structure Analysis

In the network, N is the number of nodes and K is the number of links. In weighted networks every link connecting two nodes has an associated weight $w \in R_0^+$. The co-occurrence window m_n of size n is defined as n subsequent words from a text. The number of network components is denoted by ω.

For every two connected nodes i and j the number of links lying on the shortest path between them is denoted as d_{ij}, therefore the average distance of a node i from all other nodes is:

$$d_i = \frac{\sum_j d_{ij}}{N}. \qquad (1)$$

And the average path length between every two nodes i, j is:

$$L = \sum_{i,j} \frac{d_{ij}}{N(N-1)}. \qquad (2)$$

The maximum distance results in the network diameter:

$$D = max_i d_i. \tag{3}$$

For weighted networks the clustering coefficient of a node i is defined as the geometric average of the subgraph link weights:

$$c_i = \frac{1}{k_i(k_i - 1)} \sum_{ij} (\hat{w}_{ij} \hat{w}_{ik} \hat{w}_{jk})^{1/3}, \tag{4}$$

where k_i is the degree of the node i, and the link weights $\hat{w}_{ij}$ are normalized by the maximum weight in the network $\hat{w}_{ij} = w_{ij} / \max(w)$. The value of c_i is assigned to 0 if $k_i < 2$.

The average clustering of a network is defined as the average value of the clustering coefficients of all nodes in a network:

$$C = \frac{1}{N} \sum_i c_i. \tag{5}$$

If $\omega > 1$, C is computed for the largest network component.

An important property of complex networks is degree distribution. For many real networks this distribution follows power law [13], which is defined as:

$$P(k) \sim k^{-\alpha}, \tag{6}$$

where the distribution parameter α is typically in range between 2 and 3.

4 Methodology

4.1 Data

For the construction and analysis of co-occurrence networks, we used two corpora. First is the original text of literature (C1), and second is the shuffled version of the same text (C2). In C2, the content of the original corpus is randomized by shuffling the words and punctuation marks, so C2 has the same quantity and frequency of words as the original corpus, but the text itself is meaningless.

Corpus C1 contains 10 books written in or translated into the Croatian language: I. Andrić "The Bridge on the Drina", M. Krleža "On the Edge of Reason" and "The Return of Philip Latinowicz", A. Šenoa "Branka", M. Jergović "Mama Leone", C. Collodi "Pinocchio", U. Eco "The Name of the Rose", E. Hemingway "The Old Man and the Sea", S. King "The Mist", and H. Lee "To Kill a Mockingbird".

The C1 has 895547 words, of which 91714 are unique, in 59128 sentences. The shuffling algorithm randomized words and punctuation marks which raised the new structure of sentences in the C2. The C2 has the same number of words in 58896 sentences.

4.2 The Construction of Co-occurrence Networks

Text can be represented as a complex network of linked words: each individual word is a node and interactions amongst words are links. From C1 and C2 we constructed eight different co-occurrence networks, all weighted and directed. Words are nodes linked within the co-occurrence window and according to the usage of the delimiters (punctuation marks).

The co-occurrence window m_n of size n is defined as a set of n subsequent words from a text. Within a window the links are established between the first word and $n-1$ subsequent words. In the networks where the linkage is limited to the sentence borders during the construction, we consider the sentence boundary as the window boundary too. Three steps in the network construction for a sentence of 6 words, with usage of the delimiters, for the co-occurrence window sizes $n = 2$ and $n = 6$ are shown in Fig. 1.

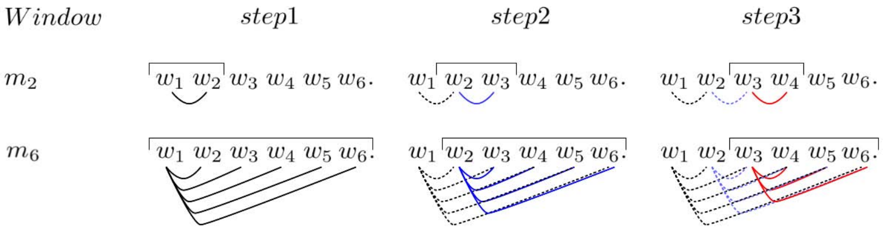

Fig. 1 An illustration of 3 steps in a network construction with a co-occurrence window m_n of sizes $n = 2$, and $n = 6$. $w_1 ... w_6$ are words within a sentence

The weight of the link is proportional to the overall co-occurrence frequencies of the corresponding words within a co-occurrence window. Network construction and analysis was implemented with the Python programming language using the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [6]. Numerical analysis of power law distributions was made with the 'powerlaw' software package [2] for the Python programming language.

5 Results

The comparison of the properties for networks differing in the co-occurrence window sizes (m_2, m_6) and the usage of delimiters are shown in Tables 1 and 2. The results show that the networks constructed with larger co-occurrence window emphasize small-world properties in both networks: from original

Table 1 Networks constructed with delimiters: the *rand* subscript is for the networks from C2

	m_2	m_6
N	91647	91647
N_{rand}	91526	91535
K	464029	2009187
K_{rand}	598519	2233643
L	3.10	2.38
L_{rand}	2.998	2.40
D	23	7
D_{rand}	9	5
C	0.32	0.71
C_{rand}	0.35	0.73
ω	22	22
ω_{rand}	15	8

Table 2 Networks constructed without delimiters: the *rand* subscript is for the networks from C2

	m_2	m_6
N	91714	91714
N_{rand}	91714	91714
K	513297	2459706
K_{rand}	636748	2666998
L	3.05	2.30
L_{rand}	2.95	2.29
D	17	6
D_{rand}	7	4
C	0.34	0.68
C_{rand}	0.38	0.70
ω	1	1
ω_{rand}	1	1

and shuffled texts. More precisely, in networks built with m_6, values of the average path length L and network diameter D are smaller, and the average clustering coefficient C is larger in comparison to the same measures from networks built with m_2.

Furthermore, in Tables 1 and 2 we compare the characteristics of networks constructed for co-occurrence window limited within or across the sentence boundaries. In the networks without delimiters, words are linked within a given co-occurrence window regardless of being in different sentences.

All of the networks constructed without the usage of delimiters show smaller network distance measures. Also, the clustering coefficient becomes larger only in the case of m_2, while the larger co-occurrence window m_6 decreases its value.

The number of nodes N ($N_{rand} < N$) is different from the number of words in C1, due to the used co-occurrence criteria. Our approach (Table 1) limits the co-occurrence window size within the sentence delimiters. This causes sentences with exactly one word to be isolated from the network, which reduces the number of nodes N. This is the reason why we considered the co-occurrence window across sentence boundaries (Table 2). $\omega_{rand} < \omega$ indicates that the number of connected components is smaller in the shuffled text C2. Therefore, when co-occurrence window disregarded the sentence boundaries networks have only 1 connected component (Table 2).

Fig. 2. shows the comparison of the plots of the clustering coefficient against the node degree for four different networks. Each plot shows clustering coefficient values spread on a log-log scale. The difference between plots constructed for networks based on original (a, b) and shuffled text (c, d) is that the clustering coefficients are more dispersed for the C1 than for the C2. It is especially emphasized in the case of small window size (m_2). The

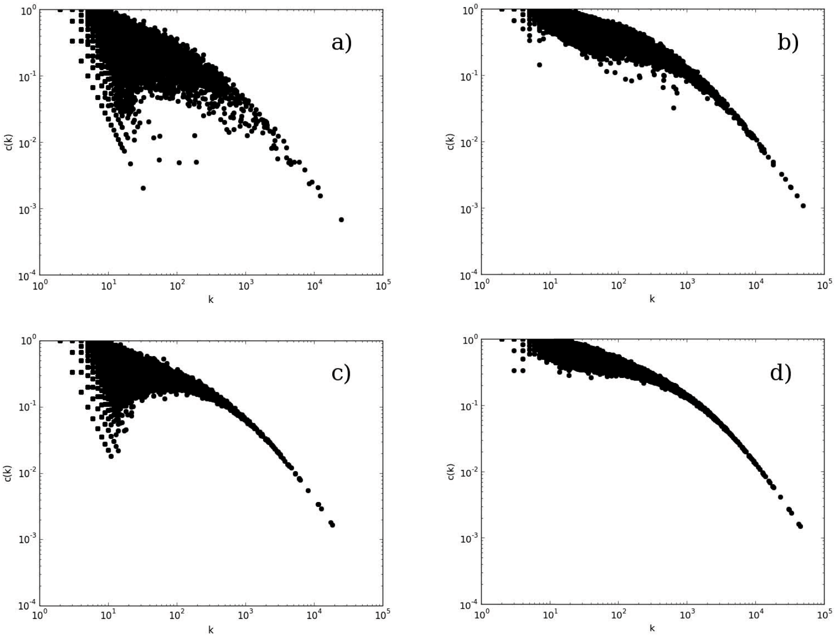

Fig. 2 Plots of the clustering coefficient against the degree of the vertices for four networks: (a) network based on the original text with m_2, (b) network based on the original text with m_6, (c) network based on the shuffled text with m_2, (d) network based on the shuffled text with m_6; always with delimiters used

dispersion of the clustering coefficient values associated with the properties of the word neighborhood reflects the complex organization of words [11]. Therefore, the more dispersed plots for the networks from the original texts, may indicate the more complex structure of original texts in comparison to the shuffled texts.

The clustering coefficient, as a local measure, is calculated considering the links' weights (Eq. 4). The results shown in Fig. 2 indicate that clustering coefficient of the weighted networks should be considered in the further study of the syntax structure.

Numerical results of power law distribution analysis indicate the presence of the power law distribution. The numeric values of α for the power law distributions are: 2.167 for m_2, C1; 2.090 for m_2, C2; 2.158 for m_6, C1; 2.137 for m_6, C2.

The global network measures: average shortest path length, diameter, clustering coefficient and degree distribution may not be well-suited properties for fine-grained network analysis. This may be explained by the fact that the syntax is a local language property. Therefore, it is necessary to include local network measures such as clustering coefficient of a node.

6 Conclusion

We studied the topologies of the linguistic networks constructed from normal and shuffled Croatian texts. As expected, the text shuffling causes the decrease of the network diameter, due to the establishment of previously nonexisting links. This indicates that syntax does play a significant role in the Croatian language, although it is a free word-order language.

We have shown that the Croatian language networks have similar properties as language networks from English and other languages. Firstly, all Croatian language co-occurrence networks, based on normal and shuffled texts, have a power law degree distribution. That means that text shuffling has no influence on the degree distribution, which has already been shown for English [11, 12], English and Portuguese [5] and English, French, Spanish and Chinese [7]. Furthermore, all eight networks constructed for the Croatian language have small-world properties. There is a slight difference in the average clustering coefficient which is higher for the networks based on shuffled text. Distance measures (average shortest path length and diameter) show that each of the four networks based on normal texts have a greater L and D value than the corresponding network based on shuffled text. The same relations for average clustering coefficient, average shortest path length and diameter are shown in [7] for all studied languages (English, French, Spanish and Chinese). Similar results are shown for English and Portuguese in [5], although the authors used different shuffling procedures.

Our results imply that the syntax structure of the Croatian language has impact on the network properties, which needs further detailed analysis in order to find which network measures perform a fine-grained differentiation between an original and shuffled text. This should be thoroughly examined in the future work, which will cover: a) the comparison of the topological properties of the networks constructed from shuffled texts with preserved sentence length frequencies, b) shuffling of each book separately, c) using the node selectivity measure, and d) the analysis of the syntax dependencies in the Croatian linguistic networks.

References

1. Meta-net white paper series: Key results and cross-language comparison (2012), `http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison`
2. Alstott, J., Bullmore, E., Plenz, D.: Powerlaw: a python package for analysis of heavy-tailed distributions. arXiv preprint arXiv:1305.0215 (2013)
3. Ban, K., Martinčić-Ipšić, S., Meštrović, A.: Initial comparison of linguistic networks measures for parallel texts. In: 5th International Conference on Information Technologies and Information Society (ITIS), pp. 97–104 (2013)
4. Barabási, A., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)

5. Caldeira, S., Lobao, P., Andrade, R., Neme, A., Miranda, V.: The network of concepts in written texts. The European Physical Journal B-Condensed Matter and Complex Systems 49(4), 523–529 (2006)
6. Hagberg, A., Swart, P., Chult, D.: Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Laboratory (LANL) (2008)
7. Krishna, M., Hassan, A., Liu, Y., Radev, D.: The effect of linguistic constraints on the large scale organization of language. arXiv preprint arXiv:1102.2831 (2011)
8. Li, W.: Random texts exhibit zipf's-law-like word frequency distribution. IEEE Transactions on Information Theory 38(6), 1842–1845 (1992)
9. Liu, H., Hu, F.: What role does syntax play in a language network? EPL (Europhysics Letters) 83(1), 18002 (2008)
10. Margan, D., Martinčić-Ipšić, S., Meštrović, A.: Preliminary report on the structure of Croatian linguistic co-occurrence networks. In: 5th International Conference on Information Technologies and Information Society (ITIS), Slovenia, pp. 89–96 (2013)
11. Masucci, A., Rodgers, G.: Network properties of written human language. Physical Review E 74(2), 026102 (2006)
12. Masucci, A., Rodgers, G.: Differences between normal and shuffled texts: structural properties of weighted networks. Advances in Complex Systems 12(1), 113–129 (2009)
13. Newman, M.: Power laws, pareto distributions and zipf's law. Contemporary Physics 46(5), 323–351 (2005)
14. Watts, D., Strogatz, S.: Collective dynamics of small-world networks. Nature 393(6684), 440–442 (1998)