Preliminary Report on the Structure of Croatian Linguistic Co-occurrence Networks

Domagoj Margan, Sanda Martinčić-Ipšić, Ana Meštrović

Department of Informatics

University of Rijeka

Radmile Matejčić 2, 51000 Rijeka, Croatia

`{dmargan, smarti, amestrovic}@uniri.hr`

Abstract. *In this article, we investigate the structure of Croatian linguistic co-occurrence networks. We examine the change of network structure properties by systematically varying the co-occurrence window sizes, the corpus sizes and removing stopwords. In a co-occurrence window of size n we establish a link between the current word and $n − 1$ subsequent words. The results point out that the increase of the co-occurrence window size is followed by a decrease in diameter, average path shortening and expectedly condensing the average clustering coefficient. The same can be noticed for the removal of the stopwords. Finally, since the size of texts is reflected in the network properties, our results suggest that the corpus influence can be reduced by increasing the co-occurrence window size.*

Keywords. complex networks, linguistic co-occurrence networks, Croatian corpus, stopwords

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. Text can be represented as a complex network of linked words: each individual word is a node and interactions amongst words are links. The interactions can be derived at different levels: structure, semantics, dependencies, etc. Commonly they rise from a simple criterion such as co-occurrence of two words within a sentence, or text.

The pioneering construction of linguistic networks was in 2001, when Ferrer i Cancho and Solé [8] showed that the co-occurrence network from the British National Corpus has a small average path length, a high clustering coefficient, and a two-regime power law degree distribution; the network exhibits small-world and scale-free properties. Drogotsev and Mendes [6] used complex networks to study language as a self-organizing network of interacting words. The co-occurrence networks were constructed by linking two neighboring words within a sentence. Masucci and Rodgers [10] investigated the network topology of Orwell's '1984' focusing on the local properties: nearest neighbors and the clustering coefficient by linking the neighboring words. Pardo *et al.* [11] used the complex network's clustering coefficient as the measure of text summarization performance. The original and summarized texts were preprocessed with stopwords' removal

and lemmatization. For the network construction they used reversed window orientation which caused the word to be connected to the previous words with forwarding links' directions. Caldiera *et al.* [4] examined the structure of the texts of individual authors. After stopword elimination and lemmatization each sentence was added to the network as a clique[1]. Biemann *et al.* [2] compared networks where two neighboring words were linked with networks where all the words co-occurring in the sentence were linked. From the network properties they derived a quantifiable measure of generative language (n-gram artificial language) regarding the semantics of natural language. Borge-Holthoefer [3] produced a methodological and formal overview of complex networks from the language research perspective. Liu and Cong [9] used complex network parameters for the classification (hierarchical clustering) of 14 languages, where Croatian was amongst 12 Slavic.

In this paper we construct the linguistic co-occurrence networks from Croatian texts. We examine the change of a network's structure properties by systematically varying the co-occurrence window sizes, the corpus sizes and stopwords' removal. In a co-occurrence window of size n we establish a link between the current word and $n-1$ subsequent words.

In Section 2 we define network properties needed to accurately analyze small-world and scale-free characteristics of co-occurrence networks, such as diameter, average path length and average clustering coefficient. In Section 3 we present the construction of 30 co-occurrence networks. The network measurements are in Section 4. In the final Section, we elaborate on the obtained results and make conclusions regarding future work.

2   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:

$$d_i = \frac{\sum_j d_{ij}}{N}. \tag{1}$$

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

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

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:

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

---

[1]A clique in an undirected network is a subset of its nodes such that every two nodes in the subset are linked.

where 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, which is defined as:

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

3 Network construction

3.1 Data

For the construction and analysis of co-occurrence networks, we used a corpus of literature, containing 10 books written in or translated into the Croatian language. For the experiments we divided the corpus into three parts: C1 - one book, C2 - four books and C3 - ten books, where C1 $\subseteq$ C2 $\subseteq$ C3, as shown in Table 1.

Stopwords are a list of the most common, short function words which do not carry strong semantic properties, but are needed for the syntax of language (pronouns, prepositions, conjunctions, abbreviations, interjections,...). The Croatian stopwords list contains 2,923 words in their inflected forms. Examples of stopwords are: 'is', 'but', 'and', 'which', 'on', 'any', 'some'.

Corpus part	C1	C2	C3
# of words	28671	252328	895547
# of unique words	9159	40221	91018
# of stopwords	371	588	629

Table 1: The statistics for the corpus of 10 books

3.2 The construction of co-occurrence networks

We constructed 30 different co-occurrence networks, weighted and directed, from the corpus in Table 1. Words are nodes, and they are linked if they are in the same sentence according to the size of the co-occurrence window. 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. During the construction we considered the sentence boundary as the window boundary too. Three steps in the network construction for a sentence of 5 words, and the co-occurrence window size $n = 2..5$ is shown in Fig. 1.

The weight of the link between two nodes is proportional to the overall co-occurrence frequencies of the corresponding words within a co-occurrence window. For all three parts of the corpus C1, C2, C3, we examined the properties of co-occurrence networks
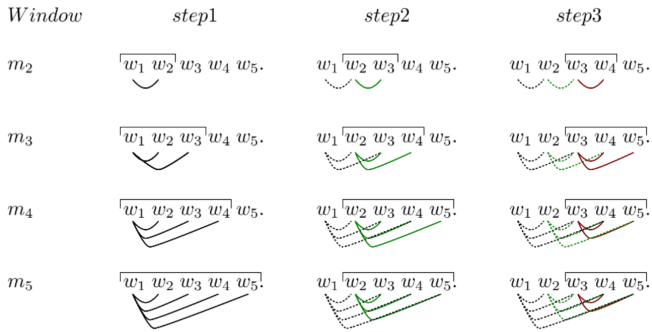
Figure 1: An illustration of 3 steps in a network construction with a co-occurrence window m_n of sizes $n = 2...5$. $w_1...w_5$ are words within a sentence.

constructed with various m_n, $n = 2, 3, 4, 5, 6$. Besides 5 window sizes for co-occurrence networks, we also differentiate upon the criterion of the inclusion or exclusion of stop-words.

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 [7]. Numerical analysis and visualization of power law distributions was made with the 'powerlaw' software package [1] for the Python programming language.

4   Results

	m_2	m_3	m_4	m_5	m_6
N_{sw}	9530	9530	9530	9530	9530
N	9159	9159	9159	9159	9159
K_{sw}	22305	43894	64161	83192	101104
K	14627	28494	41472	53596	64840
L_{sw}	3.59	2.92	2.70	2.55	2.45
L	6.42	4.73	4.12	3.79	3.58
D_{sw}	16	9	7	6	6
D	26	15	11	10	8
C_{sw}	0.15	0.55	0.63	0.66	0.68
C	0.01	0.47	0.56	0.60	0.64
ω_{sw}	5	5	5	5	5
ω	15	15	15	15	15

Table 2: Networks constructed from C1. Measures noted with the sw subscript are results with stopwords included.

The comparisons of the properties for networks differing in the co-occurrence window size are shown in Tables 2, 3 and 4. Clearly, the results show that the networks constructed

with larger co-occurrence window emphasize small-world properties. More precisely, the values of the average path length and network diameter decrease proportionally to the increase of co-occurrence window size. Likewise, the average clustering coefficient becomes larger in accordance with the increment of m_n.

	m_2	m_3	m_4	m_5	m_6
N_{sw}	40809	40809	40809	40809	40809
N	40221	40221	40221	40221	40221
K_{sw}	156857	307633	445812	572463	688484
K	108449	207437	296233	375535	446547
L_{sw}	3.25	2.81	2.64	2.52	2.43
L	4.69	3.86	3.54	3.35	3.23
D_{sw}	18	12	8	7	6
D	24	14	11	9	9
C_{sw}	0.25	0.58	0.65	0.68	0.70
C	0.02	0.43	0.52	0.56	0.59
ω_{sw}	9	9	9	9	9
ω	33	33	33	33	33

Table 3: Networks constructed from C2. Measures noted with the sw subscript are results with stopwords included.

	m_2	m_3	m_4	m_5	m_6
N_{sw}	91647	91647	91647	91647	91647
N	91018	91018	91018	91018	91018
K_{sw}	464029	911277	1315888	1680848	2009187
K	360653	684008	963078	1202869	1409599
L_{sw}	3.10	2.74	2.58	2.47	2.38
L	4.17	3.55	3.30	3.16	3.08
D_{sw}	23	13	9	7	7
D	34	19	14	11	9
C_{sw}	0.32	0.61	0.67	0.69	0.71
C	0.03	0.42	0.51	0.55	0.58
ω_{sw}	22	22	22	22	22
ω	64	64	64	64	64

Table 4: Networks constructed from C3. Measures noted with the sw subscript are results with stopwords included.

In Tables 2, 3 and 4 we also compare the characteristics of networks with the removal of the stopwords. In addition to the proportional strengthening of small-world properties with the increase of m_n, the same phenomenon appears with the inclusion of stopwords in the process of building the network. All of the networks show smaller network distance measures and greater clustering coefficient with the stopwords included.

Furthermore, stopwords have an impact on the average clustering coefficient in a way that increasing the corpus size with the stopwords included will result in a higher clustering coefficient, while increasing the corpus size with the stopwords excluded will result in a lower clustering coefficient (Fig. 2). This may be explained by the high impact of stopwords as the main hubs. Table 5 shows that stopwords are much stronger hubs than other hubs which we gain with the exclusion of stopwords.

| SW included				SW excluded			
m_2		m_6		m_2		m_6	
word	degree	word	degree	word	degree	word	degree
i (and)	29762	i (and)	67890	kad (when)	4260	kad (when)	14921
je (is)	13924	je (is)	53484	rekao (said)	2036	rekao (said)	5755
u (in)	13116	se (self)	42563	sad (now)	1494	jedan (one)	5142
se (self)	11033	u (in)	41188	reče (said)	1319	sad (now)	5062
na (on)	9084	da (yes, that)	35632	jedan (one)	1318	ljudi (people)	4836
da (yes)	8103	na (on)	29417	ima (has)	1281	dana (day)	4679
a (but)	6637	su (are)	22366	ljudi (people)	1264	ima (has)	4406
kao (as)	5452	a (but)	21919	dobro (good)	1119	reče (said)	4178
od (from)	4773	kao (as)	18141	dana (day)	998	dobro (good)	3964
za (for)	4708	ne (no)	16211	reći (say)	968	čovjek (human)	3496

Table 5: Top ten hubs in networks constructed from C3.
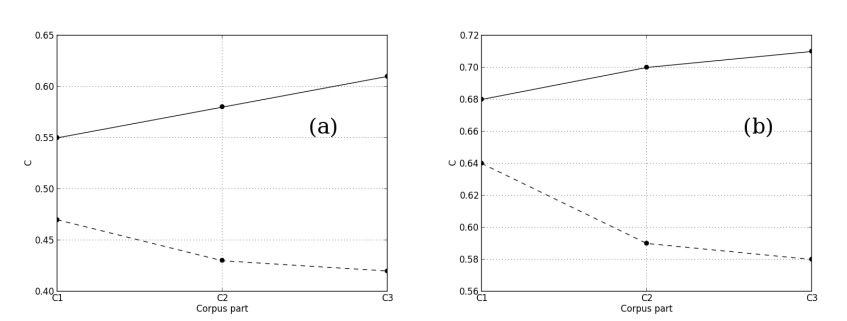
Figure 2: The impact of stopwords on the average clustering coefficient in accordance with the various sizes of the corpus parts. C_{sw} (from networks constructed with stopwords included) is represented by solid lines, while the C (from networks constructed with stopwords excluded) is represented by dashed lines. (a) m_3 networks, (b) m_6 networks.

Numerical results of power law distribution analysis indicate the presence of the power law distribution. The visualization of power law distribution for 4 networks created from C3 is shown in Fig. 3. We found that networks constructed with included stopwords generally represent a good power law fit starting from the optimal x_{min}. The numeric values of α for the power law distributions shown in Fig. 2 are respectively: 2.167, 2.172, 2.339, 2.040. The networks with stopwords included have a better power law fit.
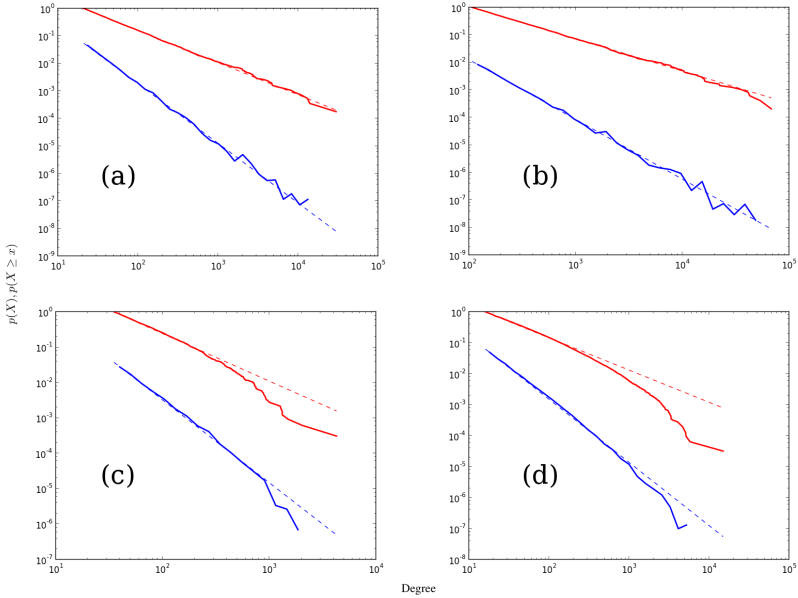
Figure 3: Comparison of plots. Probability density function ($p(X)$, lower line) and complementary cumulative distribution function ($p(X \geq x)$, upper line) of node degrees from networks constructed from C3: (a) m_2, stopwords included, (b) m_6, stopwords included, (c) m_2, stopwords excluded, (d) m_6, stopwords excluded.

5 Conclusion

In this work we have presented multiple metrics of complex networks constructed as co-occurrence networks from the Croatian language. Since, the sensitivity of the linguistic network parameters to the corpus size and stopwords [4, 5] is a known problem in the construction of linguistic networks, we analyzed the Croatian co-occurrence network. We presented the results of 30 networks constructed with the aim to examine variations among: corpus size, stopword removal and the size of the co-occurrence window.

The results in Tables 2, 3, 4, are pointing that the increase of the co-occurrence window size is followed by the diameter D decrease, average path L shortening and expectedly condensing the average clustering coefficient C. It is worth noticing, that the increased window size contributed to the results the same as the increase of the used quantity of texts did, suggesting emphasized small-world properties. The larger size of co-occurrence window plays a key role in the strengthening of properties of the small-world networks. This observation should be considered in detail in the prospect work.

Furthermore, the inclusion of stopwords in the process of network construction causes the same effect. It is evident from Table 5 that stopwords, although they have no strong semantic properties, act as hubs which can be cumbersome for semantic text analysis. The inclusion of stopwords in co-occurrence networks seems to contribute to the benefit of power law distribution, regardless of the co-occurrence window size. We point out the varying behaviour of the clustering coefficient (dynamics) by increasing the corpus size. According to our results, it depends on the presence of stopwords in the corpus: increasing

the corpus size with stopwords included, increases the value of C, while increasing the corpus size with the stopwords excluded, decreases the value of C.

Finally, since the size of texts is reflected in the network properties, our results suggest that the influence of the corpus can be reduced by increasing the co-occurrence window size. This paper is a preliminary study of the Croatian linguistic network, and more detailed research should be performed in the future. Firstly, the results should be tested on a larger corpus and power law and scale free properties proven. Additionally, the research towards extracting network semantics is a new and thrilling branch of our pursuit.

References

[1] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *arXiv preprint arXiv:1305.0215*, 2013.

[2] Chris Biemann, Stefanie Roos, and Karsten Weihe. Quantifying semantics using complex network analysis. In *COLING*, pages 263–278, 2012.

[3] Javier Borge-Holthoefer and Alex Arenas. Semantic networks: Structure and dynamics. *Entropy*, 12(5):1264–1302, 2010.

[4] Silvia MG Caldeira, TC Petit Lobao, Roberto Fernandes Silva Andrade, Alexis Neme, and JG Vivas Miranda. The network of concepts in written texts. *The European Physical Journal B-Condensed Matter and Complex Systems*, 49(4):523–529, 2006.

[5] Monojit Choudhury, Diptesh Chatterjee, and Animesh Mukherjee. Global topology of word co-occurrence networks: Beyond the two-regime power-law. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 162–170. Association for Computational Linguistics, 2010.

[6] Sergey N Dorogovtsev and José Fernando F Mendes. Language as an evolving word web. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1485):2603–2606, 2001.

[7] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Laboratory (LANL), 2008.

[8] Ramon Ferrer i Cancho and Richard V Solé. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265, 2001.

[9] HaiTao Liu and Jin Cong. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144, 2013.

[10] AP Masucci and GJ Rodgers. Network properties of written human language. *Physical Review E*, 74(2):026102, 2006.

[11] Thiago Alexandre Salgueiro Pardo, Lucas Antiqueira, M das Gracas Nunes, ON Oliveira, and Luciano da Fontoura Costa. Using complex networks for language processing: The case of summary evaluation. In *Communications, Circuits and Systems Proceedings, 2006 International Conference on*, volume 4, pages 2678–2682. IEEE, 2006.