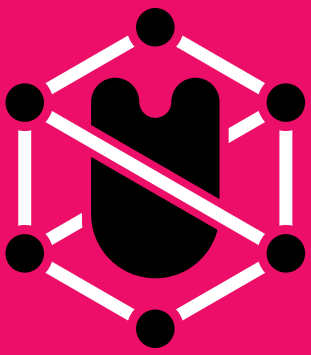


langnet

Sanda Martinčić-Ipšić
Ana Meštrović

The Language Networks

uniri

Odjel za informatiku
Sveučilišta u Rijeci

Sanda Martinčić-Ipšić
Ana Meštrović

THE LANGUAGE NETWORKS

Publisher

University of Rijeka, Department of Informatics

For the Publisher

Snježana Prijić-Samaržija

Reviewers

Prof. Dr. Matjaž Perc, University of Maribor, Ljubljana, Slovenia

Prof. Dr. Boris Podobnik, Center for Polymer Studies and Department of Physics, Boston University, Boston, USA; Faculty of Civil Engineering, University of Rijeka, Croatia; Zagreb School of Economics and Management, Zagreb, Croatia and Luxemburg School of Business, Luxemburg

Asoc. Prof. Dr. Zoran Levnajić, Faculty of Information Studies in Novo mesto, Slovenia and Jozef Stefan Institute, Slovenia

Editor

Sanda Martinčić-Ipšić

Proofread

Martin Mayhew

Logo Design

Ivo Matić

Publication Date

January 2018

Martinčić-Ipšić, Sanda and Meštrović, Ana
The Language Networks,
Includes bibliographical references and index.
1. Complex Networks 2. Natural language processing

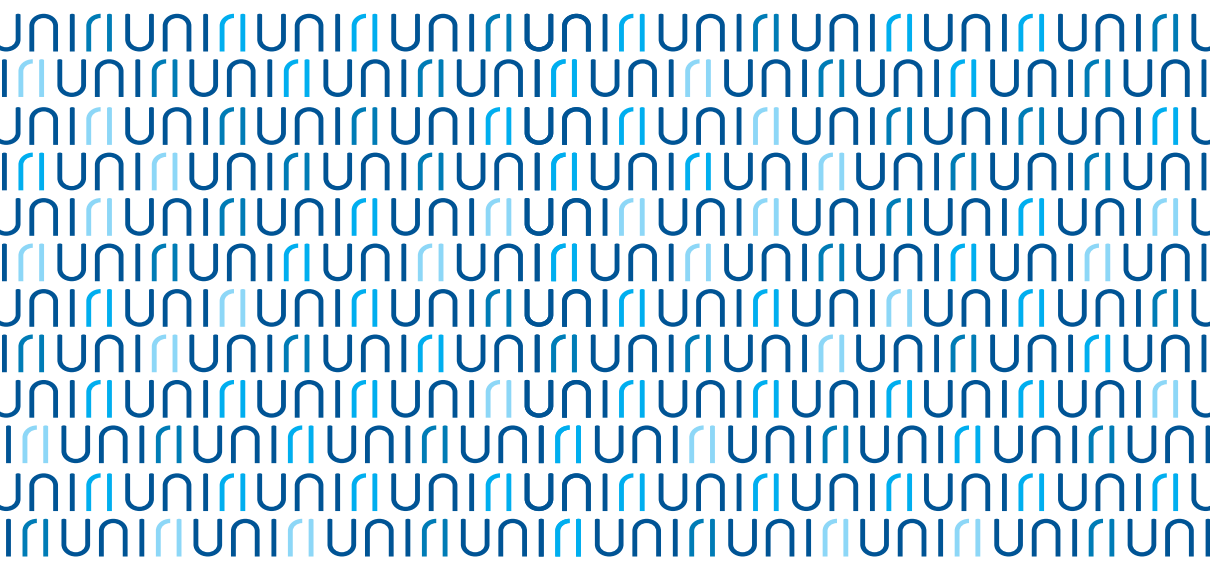
ISBN 978-953-7720-34-6

By the decision of the Publishing Committee of the University of Rijeka CLASS: 602-09/18-01/02, NUMBER: 2170-57-03-18-3 this work is published as a publication of the University of Rijeka.

langnet.uniri.hr

Sanda Martinčić-Ipšić
Ana Meštrović

The Language Networks



Rijeka, 2018.

Odjel za informatiku
Sveučilišta u Rijeci

Preface

The language networks book provides insights into the principles of modeling and analyzing structural properties of language – manly in its written form, hence text. Book guidelines the basic principles of text preprocessing, covering the very initial steps needed for any natural language processing task. Further, the book examines the possibilities of representing text in a complex networks framework. The second part overviews the application of language networks as one of a data science disciplines. It covers important data science topics for the processing of the big textual data from extracting the most salient structural parts of documents, across differentiation between text genres to predicting the speeding of the information through social media. Finally, the last part of the book is tasked with formal modeling of the linguistic subsystems in a multilayer complex networks formalism, which allows systematic study of language across all of its subsystems.

The first part of the book studies the general principles of language networks construction and analysis. It covers language network construction types. Specifically, it analyzes the effects of constructing directed vs. undirected, weighted vs. unweighted network from lemmatized (stemmed) or non-lemmatized texts with stopwords included or excluded. The effects of text randomization are studied enabling better insights into characteristics of language networks compared to their shuffled counterparts. Some preliminary experiments reveal the possibilities of the differentiation of the structural properties of networks constructed from different text types and in different languages like Croatian, English, and Italian. Next, some initial insights into the characterization of syllabic networks are presented. The analysis of motifs of the linguistic networks reveals the typical building blocks of the structure of networks of the literature in the Croatian language. Finally, the first part of the book concludes with the LaNCoA a Python Toolkit for the construction and analysis of language networks implementing the majority of the findings presented in this part of the book.

The second part of the book is dedicated to the applications of language networks. The language networks enable the extraction of the most salient words in texts – keywords and extraction of the domain knowledge-context studied on the content of Wikipedia entries. The applicative part of language networks includes the differentiation between different text types and polarization of tweets, as well. Finally, the possibilities of predicting the future content of tweets solely from the structural properties of the complex language networks are presented.

The third part of the book presents the formal model of language networks. Multilayered language network represents a comprehensive framework based on the multilayered graphs that can model various aspects of language like subsystems at the different level in the hierarchy, the construction principles, the language types and others. Multilayer language model serves as a unified formal model for the representation of language within the complex networks theory.

This book represents a collection of scientific results of the members of LangNet (Language Networks) group at Department of Informatics, the University of Rijeka from 2013. to 2017 [1]. The papers gathered in this collection have been initially published in different scientific journals and presented at scientific conferences. The complete LangNet bibliography is available at langnet.uniri.hr pages.

We are truly and sincerely thankful to many researchers that influenced this research. Foremost, Slobodan Beliga was the driving force of many research initiatives. Zoran Levnajić, Mihaela Matešić, Benedikt Perak and Tajana Ban-Kirigin initiated the constant dialog on linguist's, mathematician's and physicist's perspective on language networks which led to fruitful ideas. Finally, thanks go to all students that collaborated on LangNet research Domagoj Margan, Tanja Miličić, Neven Matas, Edvin Močibob, Kristina Ban, Hana Rizvić, Ivan Ivakić, Tomislav Bukić and Sabina Šišović. Foremost we are grateful to our families who patiently supported our work.

Finally, we express our gratitude to reviewers who contributed to the quality of content in this book. Their valuable insights improved the structure and the organization of the content.

Finally, we hope that readers of this book will gain holistic and comprehensive insights into the principles of construction and analysis of language complex networks, it's formal representation model and possible applications for different text processing tasks.

Sanda Martinčić-Ipšić and Ana Meštrović

[1] This work has been supported by the University of Rijeka under the LangNet project (13.13.2.2.07).

Contents

Language Networks Construction

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

1.1 Abstract

In this Chapter, 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.

1.2 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 [9] 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 [7] 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 [11] 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.* [12]

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 [10] used complex network parameters for the classification (hierarchical clustering) of 14 languages, where Croatian was amongst 12 Slavic.

In this Chapter we propose the construction of 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 1.3 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 1.4 we present the construction of 30 co-occurrence networks. The network measurements are in Section 1.5. In the final Section, we elaborate on the obtained results and make conclusions regarding future work.

1.3 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:

$$d_i = \frac{\sum_j d_{ij}}{N}. \tag{1.1}$$

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

$$L = \sum_{i,j} \frac{d_{ij}}{N(N-1)}. \tag{1.2}$$

The maximum distance results in the network diameter :

$$D = max_i d_i. \tag{1.3}$$

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

$$c_i = \frac{1}{k_i(k_i-1)} \sum_{ij} (\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3}, \tag{1.4}$$

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 $k_i < 2$.

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

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

$$C = \frac{1}{N}\sum_i c_i. \tag{1.5}$$

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:

$$P(k) \sim k^{-\alpha}. \tag{1.6}$$

1.4 Network Construction

1.4.1 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 1.1.

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 1.1: The statistics for the corpus of 10 books.

1.4.2 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 Figure 1.1.

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 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 stopwords.

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

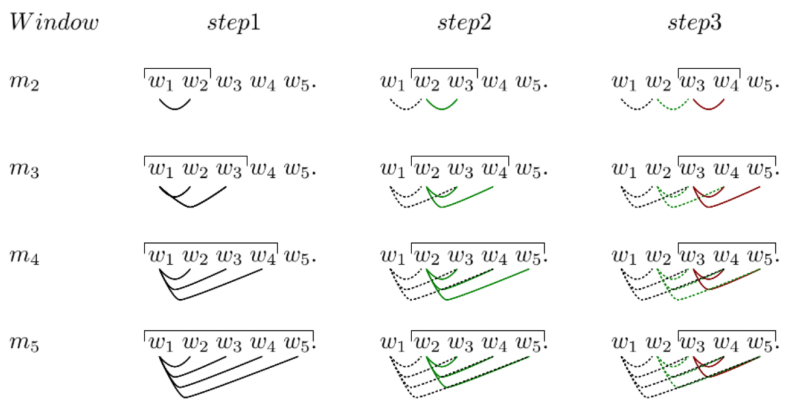


Figure 1.1: An illustration of 3 steps in a network construction with a co-occurrence window m_n of sizes $n = 2...5$. $w_1...w_5$ are words within a sentence.

1.5 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 1.2: 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 1.2, 1.3 and 1.4. 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.

In Tables 1.2, 1.3 and 1.4 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

	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 1.3: 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 1.4: Networks constructed from C3. Measures noted with the *sw* subscript are results with stopwords included.

coefficient (Figure 1.2). This may be explained by the high impact of stopwords as the main hubs. Table 1.5 shows that stopwords are much stronger hubs than other hubs which we gain with the exclusion of stopwords.

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 Figure 1.3. 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 Figure 1.2 are respectively: 2.167, 2.172, 2.339, 2.040. The networks with stopwords included have a better power law fit.

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 1.5: Top ten hubs in networks constructed from C3.

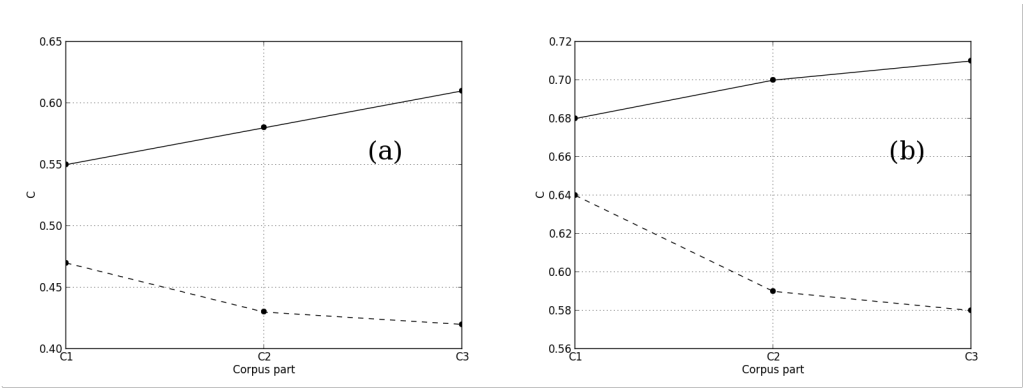


Figure 1.2: The impact of stopwords on the average clustering coefficient in accordance with the various sizes of the corpus parts. C_{sw} (from networks constructed with stopwords included) is represented by solid lines, while the C (from networks constructed with stopwords excluded) is represented by dashed lines. (a) m_3 networks, (b) m_6 networks.

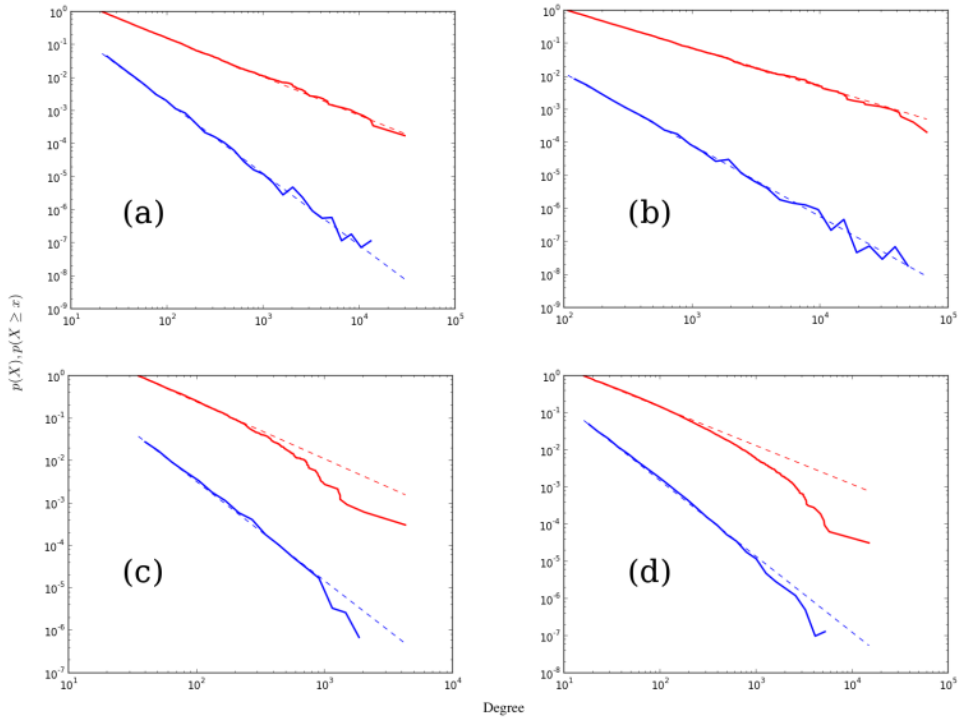


Figure 1.3: Comparison of plots. Probability density function ($p(X)$, lower line) and complementary cumulative distribution function ($p(X \geq x)$, upper line) of node degrees from networks constructed from C3: (a) m_2, stopwords included, (b) m_6, stopwords included, (c) m_2, stopwords excluded, (d) m_6, stopwords excluded.

1.6 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 1.2, 1.3, 1.4, 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 work 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.

2. Complex Networks Measures for Differentiation between Normal and Shuffled Croatian Texts

2.1 Abstract

This work studies the properties of the Croatian texts via complex networks . We present network properties of normal and shuffled Croatian texts for different shuffling principles: on the sentence level and on the text level. In both experiments we preserved the vocabulary size , word and sentence frequency distributions. Additionally, in the first shuffling approach we preserved the sentence structure of the text and the number of words per sentence. Obtained results showed that degree rank distributions exhibit no substantial deviation in shuffled networks, and strength rank distributions are preserved due to the same word frequencies. Therefore, standard approach to study the structure of linguistic co-occurrence networks showed no clear difference among the topologies of normal and shuffled texts. Finally, we showed that the in- and out- selectivity values from shuffled texts are constantly below selectivity values calculated from normal texts. Our results corroborate that the node selectivity measure can capture structural differences between original and shuffled Croatian texts.

2.2 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.

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 test whether the same properties hold for Croatian language as well. So far, there have been only sporadic efforts to model the phenomena of the Croatian language through complex networks [18] [21].

In this Chapter we address the problem of Croatian text complexity by constructing the linguistic co-occurrence networks form: a) normal texts, b) sentence-level shuffled texts, and c) text-level

shuffled texts. This work extends our previous research [21] with additional sentence-level shuffling procedure and by introducing a node selectivity as a new complex network measure. Our experiment tests whether selectivity can differentiate between normal and meaningless texts.

Section 2.3 presents an overview of related work on complex network analysis of randomized texts. In Section 2.4 we define measures for the network structure analysis. In Section 2.5 we present shuffling procedures and the construction of co-occurrence networks. The network measures are in Section 2.6. In the final Section, we elaborate the obtained results and make conclusions regarding future work.

2.3 Related Work

Some of the early work related to the analysis of random texts dates to 1992, when Li [16] showed that the distribution of words 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 [20] 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.* [4] analyzed the role played by the word frequency and sentence length distributions to the undirected co-occurrence network structure based on shuffling. Each sentence is added to the network as a complete subgraph. Shuffling procedures were conducted either on the texts or on the links. Liu and Hu [17] 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. Krishna *et al.* [15] 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. Masucci and Rodgers showed [11] [19] that the holds when they shuffled 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 proposed a measure, the node selectivity, that can distinguish a real network from a shuffled network. Selectivity is defined as the average weight distribution on the links of the single node.

Preliminary results on Croatian co-occurrence networks presented in [18] 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 [14] 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.

Initial attempt [21] to analyse network properties of normal and shuffled Croatian texts show that the text shuffling causes the decrease of the network diameter, due to the establishment of previously non-existing links. Furthermore, the results indicate a slight difference in the average clustering coefficient which is higher for the networks based on the shuffled text. Also, obtained results showed that node degree distributions are preserved in text-level shuffled networks , due to the same word frequencies (e.g. Zipf's law).

2.4 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:

$$d_i = \frac{\sum_j d_{ij}}{N}. \tag{2.1}$$

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

$$L = \sum_{i,j} \frac{d_{ij}}{N(N-1)}. \tag{2.2}$$

The maximum distance results in the network diameter :

$$D = max_i d_i. \tag{2.3}$$

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

$$c_i = \frac{1}{k_i(k_i-1)} \sum_{ij} (\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3}, \tag{2.4}$$

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:

$$C = \frac{1}{N} \sum_i c_i. \tag{2.5}$$

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

The out-degree and in-degree $k_i^{out/in}$ of node i is defined as the number of its out and in nearest neighbors.

The out-strength and the in-strength $s_i^{out/in}$ of the node i is defined as the number of its outgoing and incoming links, that is:

$$s_i^{out/in} = \sum_j w_{ij/ji}. \tag{2.6}$$

We then define for the node i the out- and in- selectivity as

$$e_i^{out/in} = \frac{s_i^{out/in}}{k_i^{out/in}}. \tag{2.7}$$

2.5 Methodology

2.5.1 Data

For the construction and analysis of co-occurrence networks, we used the corpora of 10 books written in or translated into the Croatian language. We divided the corpora into two parts: the first - C1 includes one book, and the second - C2 includes all books. The C1 contains: 191941 words (27453 unique) in 17045 sentences; and C2: 888293 words (91420 unique) in 57179 sentences.

2.5.2 The Shuffling Procedure

Commonly, the shuffling procedure randomizes the words in the text, transforming the text into the meaningless form. We shuffled the C1 and C2 corpora in two ways: a) shuffling on the sentence level, and b) shuffling on the text level . In both ways we preserved the vocabulary size, the word and sentence frequency distributions.

In the sentence-level shuffling we preserved the sentence length (the number of words per sentence) and sentence order. Versions of C1 and C2 shuffled on the sentence level are noted as C1' and C2'. In the text-level shuffling, the original text is randomized by shuffling the words and punctuation marks over the whole text. This approach preserves the number of sentences but changes the number of words per sentence. Versions of C1 and C2 shuffled on the text level are noted as C1* and C2*. Figure 2.1 shows the histograms of sentence length frequencies for C2 and C2* corpora.

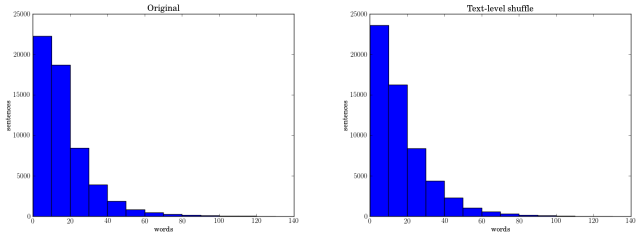


Figure 2.1: Histograms of sentence length frequencies for C2 and C2*

2.5.3 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. We constructed six different co-occurrence networks (C1, C1', C1*, C2, C2', C2*) all weighted and directed . Words are nodes linked if they are co-occurring as neighbors to each other in a sentence. The weight of the link is proportional to the overall co-occurrence frequencies of the corresponding word pairs within a corpus.

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 [8].

2.6 Results

The measures of the networks constructed from C2 corpus, sentence-level shuffled C2', and text-level shuffled C2* are presented in Table 2.1. The results show that the shuffled networks have decreased values of the average path length L and network diameter D, and have the value of the average clustering coefficient C increased.

In the networks constructed from the corpora shuffled on the text-level the number of nodes N ($N_{C1*} < N_{C1}, N_{C2*} < N_{C2}$) is smaller than the number of words in the original corpora. During the network construction process, the sentences containing just one word are disregarded, because our approach limits the word linkage to the sentence delimiters. This causes sentences with exactly one word to be isolated from the network, which reduces the number of nodes N in C1* and C2*.

One of the standard approaches to examine properties of co-occurrence networks is node degree distribution for unweighted and node strength distribution for weighted. Shuffling preserves the degree distribution due to the word frequencies [19] [21]. In this work we use rank distribution to compare different texts. Rank and frequency distributions are related through Zipf's law : the word frequency is inversely proportional to it's rank.

	C2	C2'	C2*
N	91328	91328	91204
K	465196	586110	599335
L	3.097	3.037	2.997
D	23	17	10
C	0.317	0.343	0.354
ω	22	22	10

Table 2.1: Network measures for original C2, sentence-level shuffled C2' and text-level shuffled C2*.

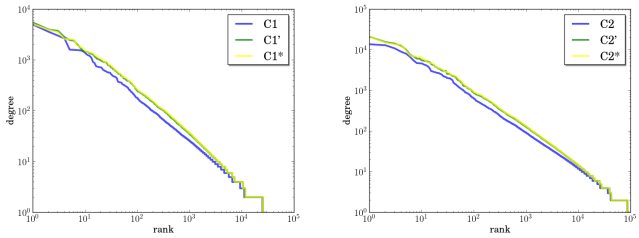


Figure 2.2: Degree rank distributions for the C1, C1', C1*, C2, C2', C2*.

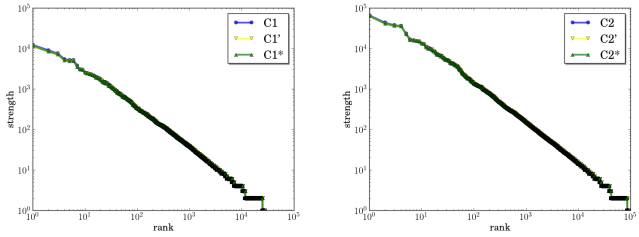


Figure 2.3: Strength rank distributions for the C1, C1', C1*, C2, C2', C2*.

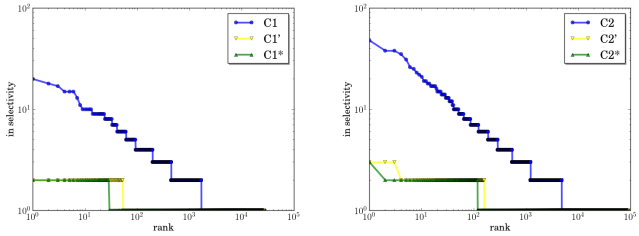


Figure 2.4: In-selectivity rank distributions for the C1, C1', C1*, C2, C2', C2*.

Initially, we computed the degree and strength rank distributions only for C2 and corresponding shuffled versions. As expected, the degree rank distributions, as well as the strength rank distributions are preserved during the shuffling procedure. In order to explore whether the same holds for substantially smaller corpus we checked degree and strength rank distributions for C1 as well. Figures 2.2 and 2.3 present degree and strength rank distributions for the C1, C1', C1*, C2, C2' and C2*. Degree rank distributions exhibit no substantial deviation in shuffled networks, and strength distribution is preserved due to the same word frequencies. At this point all network measures and distributions showed no substantial differences between the structure of the original

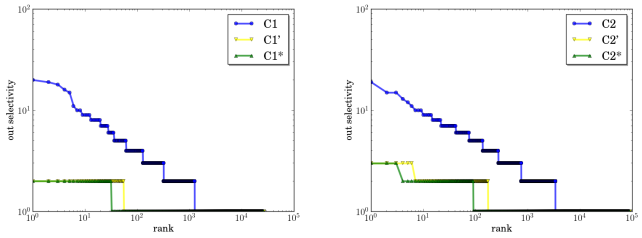


Figure 2.5: Out-selectivity rank distributions for the C1, C1', C1*, C2, C2', C2*.

and meaningless texts.

The global network measures: average shortest path length, diameter, clustering coefficient and degree and strength distribution may not be well-suited to discriminate between original and meaningless texts. Therefore, it is necessary to include local (node level) network measures. Motivated by the reported results form [19], which introduced the local measure of node selectivity, we applied the same principle on our data. In weighted and directed co-occurrence networks nodes have ingoing and outgoing links, therefore it is necessary to consider the in- and out- selectivity.

Figures 2.4 and 2.5 present obtained in- and out- selectivity ranks for C1 and C2, and their shuffled counterparts. Regardless of the links direction (in or out), selectivity values of networks from the shuffled texts are constantly below selectivity values calculated from the normal texts. This can be explained by the well known properties found in text - collocations. Collocations are typical word pairs, triples,...etc. which are recognized as standard phrases or names (e.g. New York). Shuffling procedures rearranged the word order, which disassembled collocations. This effect is reflected in the selectivity, since the high weights of common phrases and collocations are lost, which flattens and abates the selectivity ranks.

2.7 Conclusion

We studied the structure of the linguistic networks constructed from normal and shuffled texts in Croatian. As expected, the text shuffling causes the decrease of the average path length L and the network diameter D, and the increase of the average clustering coefficient C.

We showed that the degree rank distributions exhibit no substantial deviation in shuffled networks, and the strength rank distributions are preserved due to the same word frequencies. The standard approach to study the structure of linguistic co-occurrence networks showed no substantial differences among the topologies of the original and shuffled texts. Therefore we utilized the node selectivity measure proposed by Masucci and Rodgers in [19].

Our results showed that the in- and out- selectivity values from shuffled texts are constantly below selectivity values calculated from normal texts. It seems that selectivity captures typical word phrases and collocations in Croatian which are lost during the shuffling procedure. The same holds for English where Masucci and Rodgers found that selectivity somehow captures the specialized local structures in nodes' neighborhood and forms of the morphological structures in text. Based on this findings, the measure of selectivity can be useful to discriminate between different text types, which will be the part of our future work.

We have shown that the Croatian language networks have similar properties as language networks from English and other languages. Firstly, Croatian text shuffling has no influence on the degree and strength distributions, which has already been shown for English [11, 19], English and Portuguese [4] and English, French, Spanish and Chinese [15]. Furthermore, distance measures (average shortest path length and diameter) show that 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 [15] for all studied languages (English, French, Spanish and Chinese). Similar results are shown for English and Portuguese in [4], although the authors used different shuffling procedures.

Our results imply that the node selectivity is a measure suitable for fine-grained differentiation between an original and meaningless text. Furthermore, selectivity can potentially be considered as a measure for discrimination between different text types, capturing the aspect of quality of texts. This should be thoroughly examined in the future work, which will cover: a) differentiation between text types from linguistic network structure, b) finding which set of measures reflects the quality of the texts, c) the analysis of the Croatian linguistic networks using syntax dependencies instead of pure co-occurrence.

3. Comparison of Linguistic Networks Measures for Parallel Texts

3.1 Abstract

In this work we compared the properties of linguistic networks for Croatian, English and Italian languages. We constructed co-occurrence networks from parallel text corpora, consisting of the translations of five books in the three languages. We generated an Erdös-Rényi random graph with the same number of nodes and links, which enabled the comparison with linguistic co-occurrence networks, showing small-world properties. Furthermore, the comparison of Croatian, English and Italian linguistic networks showed that, besides expected commonalities of networks, there are also certain differences. The networks' measures across the three studied languages differ particularly in the shortest path length. The results indicate that size of the corpus and anomalies in text affect the network structure.

3.2 Introduction

Network analysis nowadays exhibits a growing popularity because it provides a way to analyse real complex systems. Language is an example of a complex system and in the last decade it has been the subject of many network based studies, highlighting the field of linguistic networks . Various linguistic networks can be analysed such as syntax networks [22–24], semantic networks [3], phonological networks [11, 25, 26], syllable networks [30, 31], word co-occurrence networks [9, 11].

The focus of the research in linguistic networks has shifted from one language to multiple languages. The work in [28] examines structural differences in Chinese and English by comparing the intensity and density of the connection in networks. In [27] the network properties of the English and German Wikipedia are compared. The paper by Liu and Jin [10] studied language networks on multilingual parallel texts of 15 languages. One of the 12 Slavic languages was Croatian. Network parameters were used for the hierarchical classification of the languages.

Besides multiple language studies (language differentiation and classification) the research community's attention is also focused on the genre of literature or author detection, based on the analysis of complex networks. The authors in [36] examine the correlation between the network properties and author characteristics in terms of the clustering coefficient, in and out degree, degree

distribution and component dynamics. The corpus used included over 40 books by eight authors in English. The work [32] investigates the properties of the writing style of five Persian authors in 36 books. The network derived measures: degree distribution and power law α-exponent were used for authorship identification.

Our research is an initial attempt at the analysis of parallel corpora of Croatian, Italian and English literature. We examined the comparative network properties of three languages in terms of language and book differentiation. The parallel nature of the corpus, consisting of the translations of five books in three languages, gives the opportunity to compare network properties across languages and to check the translation consistency on the book level.

Section 3.3 of the Chapter 3 presents key measures of complex networks. Section 3.4 discusses the experiments set up and in Section 3.5 the results are shown. The Chapter 3 concludes with the discussion and further research plans.

3.3 Methodology

Every network is constructed of nodes N and links K. The degree k_i of a node i is the number of connections that the node has. The average degree of the network is:

$$< k >= \frac{2K}{N}. \tag{3.1}$$

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:

$$d_i = \frac{\sum_{i \neq j} d_{ij}}{N-1}. \tag{3.2}$$

The shortest path length L is an average value of d_i of all nodes:

$$L = \sum_{i,j} \frac{d_i}{N}, \tag{3.3}$$

and the maximum distance between two nodes in the network is the diameter D:

$$D = max_i d_i. \tag{3.4}$$

The clustering coefficient c_i of a node i is described as a probability of the presence of a link between any two neighbours of a node. It is calculated as a ratio between the number of links E_i that actually exist amongst these and the total possible number:

$$c_i = \frac{2E_i}{k_i(k_i-1)} \tag{3.5}$$

The average clustering of a network C is the average value of the clustering coefficient of all the nodes:

$$C = \frac{1}{N} \sum_i c_i. \tag{3.6}$$

One of the commonly examined properties of real world networks are small-world properties [9]. The network is a small-world if its shortest path length $L \sim L_{ER}$ and its clustering coefficient $C_{ER} \gg C$ where L_{ER} is the shortest path length and C_{ER} is the clustering coefficient of an Erdös-Rényi (ER) random graph with the same number of nodes and links [39].

3.4 Experiments

3.4.1 Data

We prepared a twofold balanced corpus : parallel translations of five books in Croatian, Italian and English. Each book was originally written in one and translated to the other two languages. We took care that for each language at least one native book is present and the length of the books varies from short to long (Table 3.1).

Every text was cleared of the table of contents, the author's biography and page numbers. Afterwards the corpus was tokenized, the punctuation marks, special characters, and stopwords were removed and inflected word forms were lemmatized. For Croatian we used the stopwords list of 2922 words, for English 341 words and Italian 371 words. Table 1 shows the number of words with and without stopwords per book depending on the language. For Croatian we used the Croatian Lemmatization Server [35] for Italian and English TreeTagger [34].

Language	Book	With stopwords	Without stopwords
English	B1-EN	47684	16372
	B2-EN	147537	56525
	B3-EN	27299	10120
	B4-EN	235245	89245
	B5-EN	204517	76476
Italian	B1-IT	48487	33657
	B2-IT	156325	115855
	B3-IT	25523	20136
	B4-IT	235207	183435
	B5-IT	213147	157878
Croatian	B1-HR	44433	18627
	B2-HR	125997	59293
	B3-HR	24507	10973
	B4-HR	217987	100308
	B5-HR	198188	90299

Table 3.1: The total number of words in the books with and without stopwords by book and by language. The Croatian books show a smaller number of words but after the removal of the stopwords the total number of words is higher than in the Italian and English.

3.4.2 Networks Construction from Books

We constructed a co-occurrence network for each book: 15 directed and 15 undirected from the cleaned corpus. Words are represented as nodes and linked if they appear as adjacent words in the text. For the directed network two words are connected with and arc if one precedes the other. The same applies for the undirected network only the words are connected with a link. Additionally, we also generated an ER random graph with the same number of nodes and links for each network.

We used the Python programming language with its module NLTK [35] for text processing, the NetworkX module [8] for the construction of the networks and analysis, and Gephi software [29] for the manipulation of the networks and visualization.

3.5 Results

As shown in Tables 3.2 and 3.3, co-occurrence networks based on parallel texts share common properties: a small shortest path length L and diameter D and a high clustering coefficient C in

comparison with its associated *ER* graph. The difference between the clustering coefficient of the linguistic and the random networks varies from the minimum $C_{DIR} \approx 29C_{ER}$ to the maximum $C_{DIR} \approx 148C_{ER}$. The linguistic networks for all three languages thus have small-world properties. Another shared property of the undirected networks is a higher C and smaller L and D compared to the same measures of the directed network of the same book. This means that undirected networks are closer to the small-world networks, which is an expected result. However, there is one exception with the results for book *B*5 the diameter of which increased in the undirected network.

			Directed			Erdös-Rényi		
	N	$< k >$	C_{DIR}	L_{DIR}	D_{DIR}	C_{ER}	L_{ER}	D_{ER}
English								
B1-EN	2389	5.40	0.070	3.33	10	0.00228	4.60	15
B2-EN	7322	6.50	0.054	3.56	13	0.00089	5.34	19
B3-EN	1798	4.38	0.076	3.23	10	0.00247	4.53	14
B4-EN	12126	5.87	0.072	3.52	12	0.00049	5.93	20
B5-EN	10027	6.38	0.051	3.64	14	0.00064	5.59	20
Italian								
B1-IT	3858	4.28	0.052	3.51	13	0.00111	5.06	17
B2-IT	9120	6.45	0.044	3.64	13	0.00071	5.51	21
B3-IT	2269	4.30	0.068	3.32	10	0.00191	4.65	14
B4-IT	14009	6.34	0.047	3.62	14	0.00045	5.95	22
B5-IT	13403	5.86	0.044	3.65	14	0.00044	6.01	20
Croatian								
B1-HR	4155	3.74	0.047	3.65	12	0.00090	5.09	16
B2-HR	12610	4.23	0.034	3.92	13	0.00033	5.93	21
B3-HR	2970	3.23	0.049	3.51	11	0.00110	4.69	15
B4-HR	15256	5.40	0.051	3.74	13	0.00036	6.20	20
B5-HR	15985	4.91	0.038	3.87	14	0.00031	6.25	21

Table 3.2: The results for the **directed** networks of five books in three languages: N number of nodes, $< k >$ average node degree, C_{DIR} clustering coefficient, L_{DIR} shortest path length, D_{DIR}. C_{ER} clustering coefficient, L_{ER} shortest path length and D_{ER} diameter of ER random graph.

Further analysis of B5 revealed a proportion of Latin and German, where Latin names, were inflected in Croatian, and subsequently not lemmatized, which caused additional anomalies in the results. The English lemmatizer failed due to the same problem too.

Table 3.4 presents network measures for the B5 after the removal of Latin and German words. Compared to the initial B5 results from Tables 3.2 and 3.3 the D_{DIR} and D_{UNDIR} has decreased as expected. The undirected network had changed more than the directed. The results suggest that the Latin and German parts from the book created loops which caused C_{DIR} to decrease. At the same time B5 in Italian behaves differently due to the close nature of Italian and Latin, which was partially captured during lemmatization.

The differences across languages are presented in Figure 3.1: in general, English has a higher clustering coefficient C than Croatian.

			Undirected			Erdös-Rényi		
	N	$< k >$	C_{UNDIR}	L_{UNDIR}	D_{UNDIR}	C_{ER}	L_{ER}	D_{ER}
English								
B1-EN	2389	10.8	0.145	3.32	8	0.005	3.52	6
B2-EN	7322	13	0.109	3.36	8	0.002	3.74	6
B3-EN	1798	8.76	0.155	3.30	8	0.004	3.67	6
B4-EN	12126	11.74	0.144	3.52	8	0.001	4.07	7
B5-EN	10027	12.76	0.103	3.60	23	0.001	4.00	7
Italian								
B1-IT	3858	8.56	0.108	3.45	9	0.003	4.08	7
B2-IT	9120	12.9	0.088	3.35	11	0.001	3.83	6
B3-IT	2269	8.6	0.137	3.29	9	0.004	3.82	7
B4-IT	14009	12.68	0.096	3.42	9	0.001	4.02	6
B5-IT	13403	11.72	0.088	3.60	19	0.001	4.12	7
Croatian								
B1-HR	4155	7.48	0.099	3.58	10	0.002	4.36	8
B2-HR	12610	8.46	0.069	3.67	11	0.001	4.67	8
B3-HR	2970	6.46	0.098	3.54	9	0.003	4.47	8
B4-HR	15256	10.8	0.103	3.49	10	0.001	4.31	7
B5-HR	15985	9.82	0.077	3.77	22	0.001	4.49	8

Table 3.3: The results for the **undirected** networks of five books in three languages: N number of nodes, $< k >$ average node degree, C_{UNDIR} clustering coefficient, L_{UNDIR} shortest path length, D_{UNDIR}. C_{ER} clustering coefficient, L_{ER} shortest path length and D_{ER} diameter of ER random graph.

	N	$< k >$	C_{DIR}	L_{DIR}	D_{DIR}
B5-EN	9355	6.754	0.054	3.59	13
B5-IT	10674	6.739	0.051	3.53	13
B5-HR	12817	5.463	0.042	3.82	14
	N	$< k >$	C_{UNDIR}	L_{UNDIR}	D_{UNDIR}
B5-EN	9355	62754	0.108	3.42	17
B5-IT	10674	6.739	0.103	3.43	15
B5-HR	12817	5.463	0.085	3.54	15

Table 3.4: The new values for the directed *DIR* and undirected *UNDIR* networks of B5 by language.

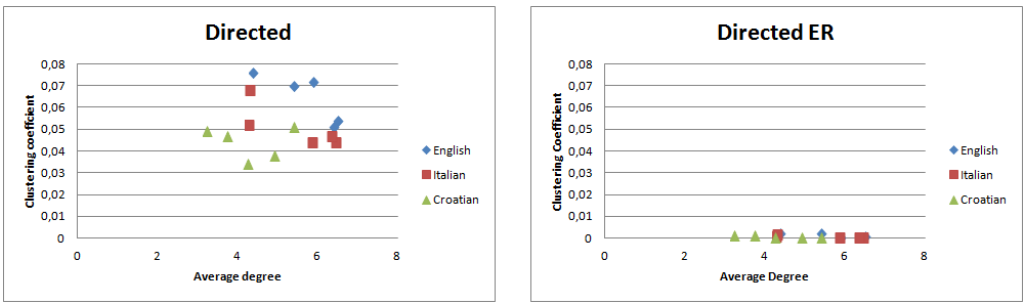


Figure 3.1: Values of average degree and clustering coefficient for 15 directed and 15 ER random networks grouped in languages.

Shortest path lengths L are the highest for Croatian, in the middle for Italian and the lowest for the English language networks as shown in Figure 3.2. Similar results are presented in [8] where it is shown that Croatian language has larger values of L and D but C twice as small than those of English. According to the graphs shown in Figure 3.2 the shortest path length seems to be more influenced by the language than diameter. D depends on the book size, but it is also sensitive to potential anomalies in the book's language, as is previously shown for book 5.

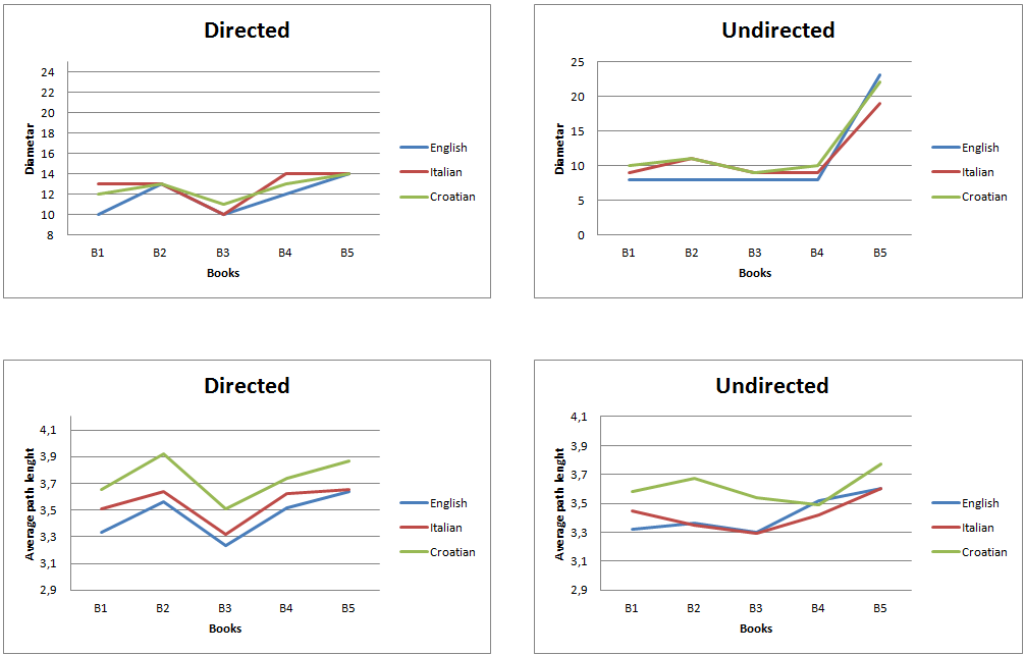


Figure 3.2: In the first row the ratio between the diameter of the books by language for directed and undirected networks is shown. The second row is the differentiation by the shortest path length.

3.6 Conclusion

In this Chapter we have examined linguistic networks for Croatian, English and Italian language. The measures of 30 directed and undirected co-occurrence networks for five books in three languages have been compared.

It has been shown that for all languages co-occurrence networks share small-world properties and corpus-sensitivity. Corpus size and possible anomalies in the text have an impact on the network structure in all three languages. An anomaly, such as the introduction of another language causes that the diameter of an undirected network becomes much higher than the diameter of a directed network as has been shown in the case of book B5. In addition, the results show that there are expected differences between the measures for directed and undirected networks for all three languages.

However, further examination of the measures of networks differs across languages: the clustering coefficient of English and Italian books is closer than that of Croatian. The Croatian language exhibits a higher path length in both directed and undirected networks, which can be caused by the relatively free word order. The word order of English is more precise than the Italian which is reflected in the directed networks in Figure 3.2. The Croatian language also has the smallest clustering coefficient which can indicate a richer language morphology. This result is partly sensitive to the degraded lemmatization of Croatian, which is also grounded in its complex morphology.

Finally, the shortest path length and clustering coefficient show language differentiation potential and should be analysed on larger corpora to test if they may be used as language classifiers. On the other hand the diameter is more related to books, which implies that it could be used as measure of the authors' vocabulary or verbosity. In further work all results should be tested on larger corpora in more languages in order to classify authorial or book genres from network parameters.

4. A Preliminary Study of Croatian Language Syllable Networks

4.1 Abstract

This research presents preliminary results of Croatian syllable networks analysis. Syllable network is a network in which nodes are syllables and links between them are constructed according to their connections within words. In this work we analyze networks of syllables generated from texts collected from the Croatian Wikipedia and Blogs. As a main tool we use complex network analysis methods which provide mechanisms that can reveal new patterns in a language structure. We aim to show that syllable networks have much higher clustering coefficient in comparison to Erdös-Renyi random networks . Furthermore, our results have been compared with other studies on syllable networks of Portuguese and Chinese and we show that Croatian syllable networks have similar properties as Portuguese and Chinese syllable networks. The results indicate that Croatian syllable networks exhibit certain properties of a small world networks.

4.2 Introduction

Network analysis has become significant method in different research areas such as biology, computer science, economics, sociology, medicine and linguistics. Complex networks are a class of networks that exhibit specific topological features, such as high clustering coefficients, small diameters, power-law degree distribution, community structures, one or several giant components, hierarchical structures, etc. Two important classes of complex networks that can be further differentiated are small-world networks [20, 40, 41] with high clustering as a main property and scale-free networks [42,43] which can be characterized by power-law degree distribution. Language can be viewed as a complex network if it is presented as a system of interacting units. Network analysis provides mechanisms that can reveal new patterns in a complex structure and can thus be applied to the study of the patterns in language structures. This, in turn, may contribute to a better understanding of the organization and the structure and evolution of a language.

Network properties of written human languages have already been analyzed in different research studies [3]. Networks based on co-occurrence of words in sentences are analyzed in [7, 9, 22].

The topology of human written language, through a network representation of Orwell's 1984, is presented in [11], while the co-occurrence properties of words in different languages are studied in [5]. All these studies have shown that language networks exhibit properties indicative of small-world networks, e. g. Pemble and Bingol [27] have constructed two complex networks out of Wikipedia English and German corpora and analyze conceptual networks in different languages.

So far, syllable networks have been constructed exclusively for Portuguese [30] and Chinese [44]. In both experiments, syllable networks have a large clustering coefficient and power-law degree distribution, as opposed to the Erdös-Rényi (ER) random networks [45], which have low clustering coefficient and Poisson-like degree distribution. In [30] the syllable network is used to demonstrate that language in itself resembles a living organism, evolving in time and space.

Network analysis of the syllables connections in the words may be of theoretical interest in the domain of phonology, morphology and language topology [46]. Analyses of properties of syllable networks can help in determining the phonetic structure of a language, as well as providing necessary grounds for further linguistic research. Besides theoretical analysis of language, syllable network analysis may be of certain interest in the domain of natural language processing , for speech recognition and speech synthesis. Syllables can be used as acoustic units in automatic speech recognition and as units in text-to-speech systems [47–49]. In [48, 49] a syllable-based language model is presented and it corresponds to the weighted syllable network.

In this Chapter we describe experiments with syllable networks for the Croatian language. We constructed four different syllable networks from texts collected Croatian Wikipedia and Blogs. The main goal was to analyze if the Croatian language syllable networks have properties of small-world networks and to analyze if these properties are similar to the properties of Portuguese and Chinese syllable networks. Furthermore, the aim was to compare two different strategies for network construction. As well, we wanted to compare networks from two different text corpora. The presented work is the first attempt to model Croatian syllables as the complex network.

In the Section 4.3 we present different syllable network construction strategies, text corpora and syllable networks that are constructed from the text. In the Section 4.4 we describe how to estimate network measures. In the Section 4.5 we present results. In the Section 4.6 we elaborate on the obtained data and provide concluding remarks.

4.3 Networks Construction

4.3.1 Syllable Networks Construction Strategies

Different strategies can be applied in building syllable networks from text. The idea of a syllable network is to represent syllables as nodes and establish links between them according to their connections within words. Generally speaking, a syllable network can be either undirected or directed and unweighted or weighted . In a directed syllable network, a directed link indicates the direction of the connection; displaying which syllable (node) is the initial and which syllable (node) is the target. By using a directed network, the successor or the predecessor of an intended syllable can be seen, possibly providing the grounds for further statistical analysis of language structure on the phonetic level . Weighted syllable networks contain information about the number of established links between two syllables, which is again significant in phonetic structure analysis.

A question of how to establish the links between the nodes (syllables) must be discussed. One way is to connect the syllables that belong to the same word (syllable co-occurrence network) and another way is to connect only the neighbour syllables (first-neighbour network). This results in eight different syllable network models. In [30, 44], the network is constructed in a way that two nodes (syllables) are connected if they belong to the same word, making the network undirected and unweighted. This simplified model of a syllable network is constructed in order to study the evolution of the language using phonetic elements [30]. We constructed three networks according

to this model. In our opinion, for some purposes that include natural language processing and linguistic studies, it also makes sense to construct a syllable network of syllables that are direct neighbours in the word. Therefore we additionally constructed and analyzed one directed and weighted neighbour network.

4.3.2 Data

We analyzed different networks of syllables from different text corpora . The texts used for building the networks are two large corpora. The first corpus is the Croatian Wikipedia . The second corpus contains 3.218 articles collected from different Croatian blogs (including 4 religious and 5 political portals, 6 blog spaces, 3 web-pages with comments and 4 columns from the daily newspapers).

The reason why we have chosen these corpora is because our future work is focused on the text collected from the Web. Another possible approach is to choose a dictionary of Croatian language as a network source. But in [30] it is shown that there is no big difference between syllable networks constructed from the book and syllable network constructed from the dictionary for unweighted networks. A problem we encountered was the Wikipedia corpus containing a certain number of foreign words. This is the reason why the initial network had certain syllables unusual for the Croatian language. Therefore, we examined a filtered network from which all nodes with small degree (meaning that they contain some rare and unusual syllables) were excluded. There is a linguistic difference between the two corpora. The Wikipedia corpus is more formal, so there are more standard words with a pattern in writing. On the other hand, the blog corpus is mostly written in an informal manner, with the use of dialect, slang or abbreviations. However, all of the mentioned texts specifics collected from the web are essential for our future work.

4.3.3 Syllable Networks

We constructed four different networks. Three of them were designed as word co-occurrence syllable networks: the first from the Wikipedia text - C_W, the second from the blog text - C_B, and the third was devised from both corpora - C_{WB}. The fourth network was constructed as a directed and weighted first-neighbour syllable network from the Wikipedia text. The number of nodes and edges for all four networks are displayed in Table 4.1.

	C_W	C_B	C_{WB}	$C_W - Dir$
Nodes (N)	4284	2000	4067	4438
Links (K)	170248	36202	173660	3334

Table 4.1: The number of nodes and edges in the four syllable networks.

Network construction is implemented in the Python programming language which contains the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [8]. For network visualization we used Gephi software [29]. For the separation of a word into syllables we use syllabification algorithm that is implemented according to the rules described in [46]. The syllabification process is Both corpora where in txt file format which made the reading and processing easy and the only problem was the encoding because of our diacritical signs such as č, ć, š, ž etc. The NetworkX module provided us with all the necessary commands to construct a graph and then export it in the desired format. The co-occurrence syllable network constructed from texts from Wikipedia (C_W) visualized using Gephi is shown in Figure 4.1. The most frequent syllables are pointed out.

Another co-occurrence syllable network constructed from blog corpus (C_B) is shown in Figure 4.2. This is a smaller network with smaller number of nodes, but the most frequent nodes (syllables) are similar to the first network, which is discussed in the Section 4.4.

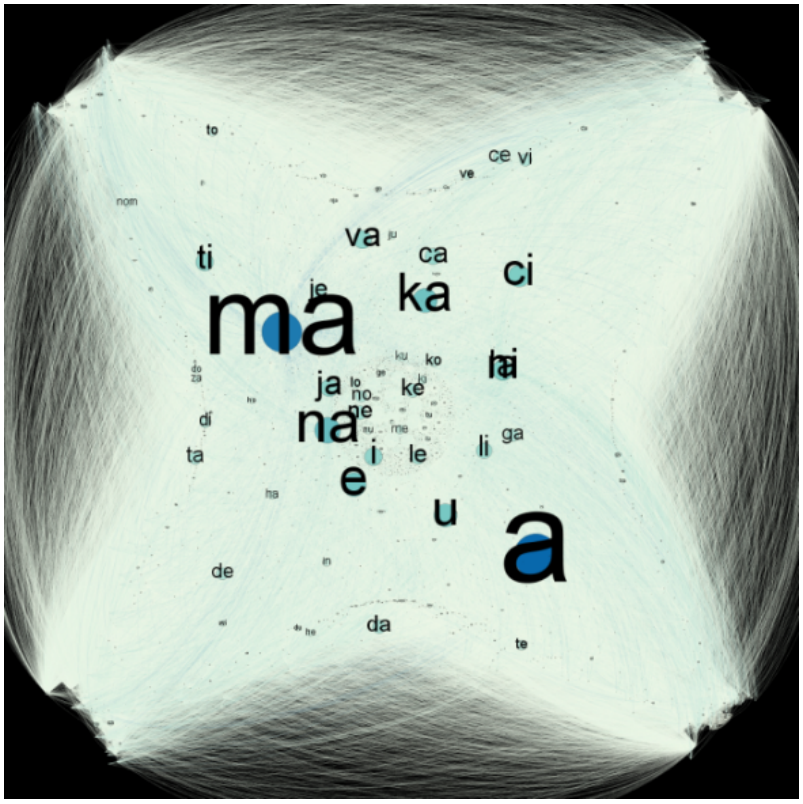


Figure 4.1: Syllable network from Wikipedia C_W.

The third network is constructed from both Wikipedia and blog corpora . Syllables with the most connections with other syllables are pointed out and are almost the same as in the first network. The fourth network is constructed as a directed and weighted network of first neighbor syllables from words that appear in the texts of the Croatian Wikipedia. The idea was to compare this network to the other three networks and to see if it had potential in phonetic structure analysis.

4.4 The Network Structure Analysis

In this Section we explain the most important measures for network analysis. Every network has a number of nodes N and links K. The degree of a node i is the number of connections of the node and is denoted by k_i. Thus, the average degree of the network is:

$$< k >= \frac{2K}{N}. \tag{4.1}$$

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:

$$d_i = \frac{\sum_j d_{ij}}{N}. \tag{4.2}$$

From where we easily obtain the average path distance L as the average value of d_i of all nodes:

$$L = \sum_{i,j} \frac{d_{ij}}{N(N-1)}, \tag{4.3}$$

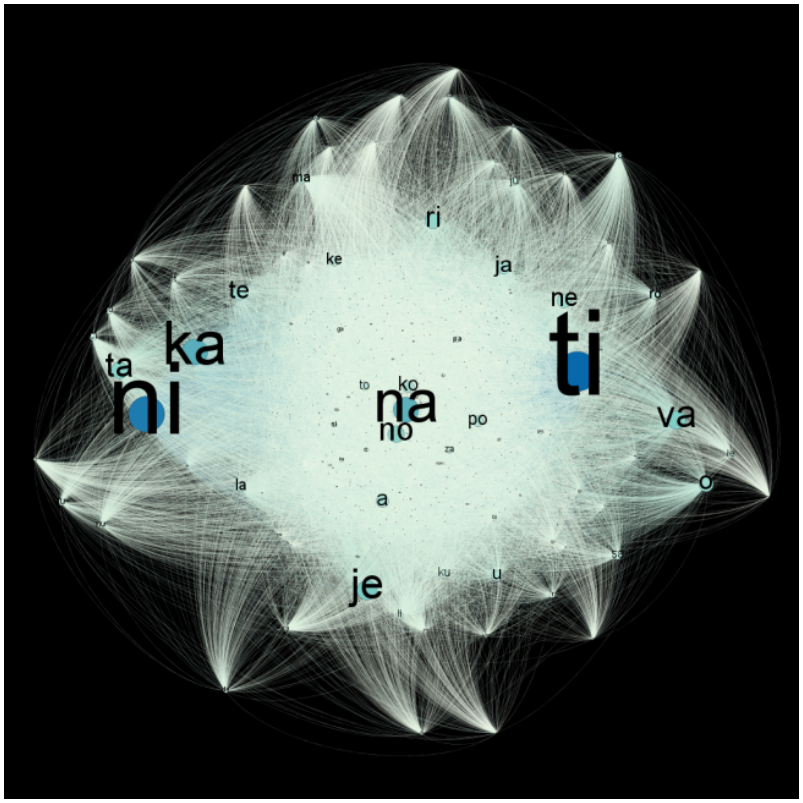


Figure 4.2: Syllable network from blog corpus C_B.

and the maximum distance results in the network diameter D:

$$D = max_i d_i. \tag{4.4}$$

The clustering coefficient is described as a presence of connections between the nearest neighbours of a node. The clustering coefficient C_i of a node i is defined as a ratio between the number of edges E_i that actually exist among these k_i and the total possible number of edges:

$$C_i = \frac{2E}{K(K-1)}. \tag{4.5}$$

The average clustering of a network C is the average value of the clustering coefficient of all the nodes:

$$C = \frac{\sum_i C_i}{N}. \tag{4.6}$$

The main property of small-world networks is that the distance between two random nodes grows proportionally to the logarithm of the number of nodes. Therefore, small-world networks tend to have small diameter and short average distance which is the property of random ER networks. Another important property is the high clustering coefficient in comparison to random ER networks. Furthermore, for complex networks it's typically a power-law degree distribution .

4.5 Results

One of our objectives in this experiment is to see if constructed syllable networks of the Croatian language have properties of small-world networks. Small-world properties have already been proven

for syllable networks of the Portuguese and the Chinese language; therefore we expected to find similar results for the Croatian language. For the purpose of comparing constructed networks with random networks, ER networks with the same number of nodes and edges have been constructed and all the important properties have been analyzed. Using Gephi we filtered the networks and determined the average degree $< k >$, diameter D, average distance L, average clustering coefficient C and some other network values. The correspondent values of these coefficients are shown in Table 4.2.

	C_W	ER_W	C_B	ER_B	C_{WB}	ER_{WB}
N	4284	4284	2000	2000	4067	4067
$< k >$	39.74	39.74	18.1	18.1	42.7	42.7
D	4	3	4	3	3	3
L	2.151	2.209	2.310	2.489	2.113	2,143
C	0.691	0.017	0.687	0.016	0.690	0.021

Table 4.2: Estimated network measures for co-occurrence syllable networks.

The results show that all three co-occurrence syllable networks have a small diameter and average path distance. Furthermore, for all three networks it holds $< k > \ll N$ which shows that syllable networks are sparse as it is expected for complex networks. In comparison to the ER networks with the same number of nodes and edges these networks show a high clustering coefficient: $C(C_W) \approx 40C(ER_W); C(C_B) \approx 42C(ER_B); C(C_{WB}) \approx 33C(ER_{WB})$. All these results lead to a conclusion that co-occurrence syllable networks of Croatian language exhibit small world network properties. We compared our results with the results obtained for the Portuguese and Chinese languages and concluded that there is a similarity between these families of syllable networks. Syllable networks of the Portuguese and of the Croatian language are similar in size, are both sparse, have a small diameter, small size of average path length and both have a high clustering coefficient. Syllable networks of the Chinese have different sizes, but the properties show that these are also small-world networks.

The results of the fourth, weighted and directed first-neighbour syllable network analysis are shown in Table 3. Although C for the first-neighbour syllable network was smaller than in the co-occurrence syllable networks, in comparison with the random network [1], it was still about 30 times larger than random network C. These values indicate that the first-neighbour syllable network may be a small-world network as well, however, more experiments with larger corpora need to be conducted.

	$C_W - Dir$	$C_W - Undir$	ER
N	4438	4438	4438
K	33341	33341	33341
D	9	8	5
C	0.153	0.208	0.007

Table 4.3: Estimated network measures for the first-neighbour syllable network.

In these preliminary experiments we did not estimate the degree distributions for syllable networks. However, we did use NetworkX functions to plot degree distributions on the log-log scale and the result that we got for the co-occurrence syllable network is shown in Figure 4.3. The

[1] For the purpose of comparison to the random network, it was transformed into an undirected, unweighted network.

straight line on log-log scale indicates that a power-law distribution should be tested in further experiments.

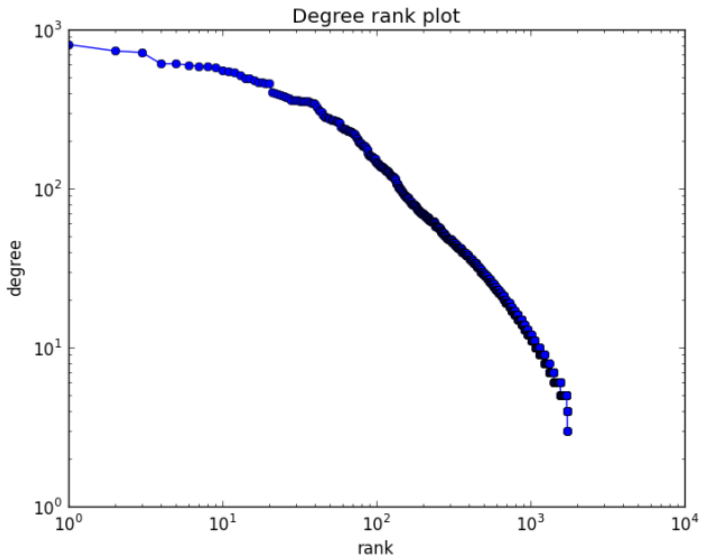


Figure 4.3: Degree distribution for co-occurrence syllable network.

Small subnetworks were filtered with ten nodes each from all three networks with the highest degree. The results are shown in Table 4.4. It is shown that all three networks have almost the same nodes with the highest degree. This indicates that different corpora do not create significant differences between the networks.

C_A		C_W		C_B	
Syll.	Degree	Syll.	Degree	Syll.	Degree
ma	2299	ma	2296	ma	836
na	2166	na	2156	ti	824
ni	1937	ni	1927	na	753
ti	1918	ra	1897	ni	741
ra	1894	a	1890	ka	627
a	1860	ti	1889	ci	626
ne	1808	ne	1792	ra	623
ka	1801	ka	1773	je	604
o	1692	o	1672	no	595
ta	1682	ta	1670	ne	593

Table 4.4: The most frequent syllables.

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 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 stopwords.

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

4.6 Conclusion

In this Chapter we presented different approaches in syllable networks construction. Undirected and unweighted word co-ccurrence syllable networks have been already constructed and analyzed for two languages: Portuguese and Chinese. The same syllable networks constructed for the Croatian language (from different corpora) exhibited similar results. The networks contain a high cluster coefficient compared to random networks of the same size and small diameter and average path length. In conclusion, the Croatian language syllable networks have properties of small-world networks. Another approach was to construct a directed and weighted first-neighbour syllable network for the Croatian language. As far as we know, this is the first time this syllable network construction type has been utilized. The main idea of this approach is to capture more information about the properties of each syllable (the successor, the predecessor and strength of connections with other syllables). It is shown that this kind of network has small-world network properties as well. These are just preliminary results and there is still a lot of future research to be conducted in this direction. The syllabification algorithm from [46] should be reconsidered and the correctness of the Croatian syllabification should be assessed. Furthermore, detailed statistical analysis should be performed. The experiment should be repeated with larger corpora such as Croatian literature and dictionaries. However, it is necessary to determine an exact degree distribution for all networks. Our plan is to analyze the network growth and possible communities in the network.

5. Network Motifs Analysis of Croatian Literature

5.1 Abstract

In this research we analyse network motifs in the co-occurrence directed networks constructed from five different texts (four books and one portal) in the Croatian language. After preparing the data and network construction, we perform the network motif analysis. We analyse the motif frequencies and Z-scores in the five networks. We present the triad significance profile for five datasets. Furthermore, we compare our results with the existing results for the linguistic networks. Firstly, we show that the triad significance profile for the Croatian language is very similar with the other languages and all the networks belong to the same family of networks. However, there are certain differences between the Croatian language and other analysed languages. We conclude that this is due to the free word-order of the Croatian language.

5.2 Introduction

Many scientists from different disciplines study networks because of their ubiquity. The complex networks in nature share global properties such as small-world property of short paths between vertices and highly clustered connections [55]. In addition, many of these networks are scale-free networks, characterised by power-law degree distribution [43]. However, besides these global network characteristics, there are certain properties on the meso-scale and local-scale [3] that explain structural differences between complex networks. That is why more detailed network analysis on the meso-scale and on the local-level is important. Network analysis on the meso-scale and local-scale may include: community detection [13], motif analysis [51] or graphlet analysis [53].

In this resarch we are focused on the network motifs' analysis. Network motifs are connected and directed subgraphs occurring in complex networks at numbers that are significantly higher than those in randomized networks [51]. Motifs may contain up to 8 vertices. For now, there have been reports on 3-vertex and 4-vertex motifs due to the complexity of the algorithm that identifies the motifs from the complex networks.

Alon et al. [52] analyse superfamilies of networks based on significant motifs (Figure 5.1). The first group of networks are from three microorganisms: the Escherichia coli, Bacillus subtilis and the Saccharomyces cerevisiae. These microorganisms form sensory transcription networks, the vertices represent genes or operons and the edges represent direct transcriptional regulation. They form the first superfamily which includes three types of biological networks: signal-transduction interactions in mammalian cells, developmental transcription networks arising from the review of the development of the fruit fly and sea-urchin, and synaptic wiring between neurons in Caenorhabditis elegans. They also studied three WWW networks of hyperlinks between web pages related to university, literature and music. A feature of these networks is the transitivity of the relations, as evidenced by the motifs presented in these networks that are highly transitive. Similar results are obtained by testing three social networks, where people from the group are represented by vertices. The connections between two people, a positive opinion of one member of the group to another member, were represented by edges, obtained on the basis of questionnaires. The conclusion is that social networks and the web are probably members of the same superfamily, which may facilitate the understanding of the structure of the web. Furthermore, word-adjacency networks are analysed so that each vertex represented a single word, and each edge represented a connection between the two words that have followed one another in the text. The results obtained for different texts in different languages (English, French, Spanish and Japanese) are similar. Significant triads are from ID3#1 to ID3#6 (considering the IDs in [52]), and underrepresented are all other triads, from the ID3#7 to ID3#13. This means that the examined languages do not have a transitive relation such as the WWW. The explanation for these results may be in the structure of language, where words are divided into categories and generally the rule is that a word from one category follows a word from the other category. As an example, most connected category words are prepositions and behind them usually follows a noun or an article. Biemann at al. [2] use motifs to quantify the differences between a natural and a generated language. The frequencies of three-vertex and four-vertex motifs for six languages are compared with artificially generated language from n-grams. An n-gram is contiguous sequence of n units (words) reflecting the statistical properties of a given text (or speech). The authors show that the four-vertex motifs can be interpreted by semantic relations of polysemy and synonymy of words.

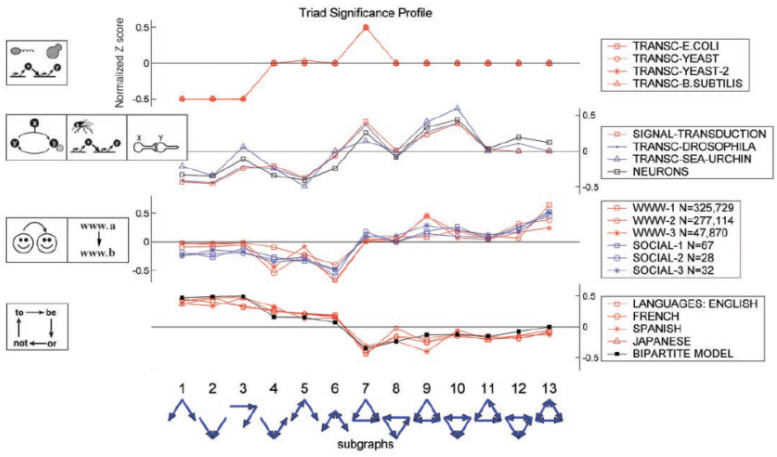


Figure 5.1: Superfamilies of complex networks according to the triad significance profile [52].

Our motivation for this research was to determine whether the local structure of the Croatian language networks share the same properties as other language networks. Croatian is a highly flective Slavic language and words can have seven different cases for singular and seven for plural, genders and numbers. The Croatian word order is mostly free, especially in non-formal writing.

These features position Croatian among morphologically rich and free word-order languages. So far Croatian has been quantified in a complex networks framework based on the word co-occurrences [14, 18] and compared with shuffled counterparts [21, 50].

In this Chapter we describe the network motifs analysis of the co-occurrence directed networks constructed from the Croatian texts: four books and one forum. We use an approach based on the significance profile (*SP*) presented in [51]. We analyse three-vertex subgraphs called triads and present the results of triad significance profile (TSP) for the five analysed networks and we compare our results with TSP for other languages.

In the Section 5.3 we give an overview of network motifs. In the Section 5.4 we describe the experiment, and the Section 5.5. presents the results. We conclude with some finishing remarks and the plans for future work.

5.3 Network Motifs

A network motif is a small subgraph that appears more frequently in the real network than in the random network. The motif may be referred to as a significantly overrepresented subgraph in the network. As well, an underrepresented subgraph in the network is called an anti-motif. In [51] authors define network motifs as small patterns for which the probability of occurrence in a randomized network is less than the probability of occurrence in the real network with the cut-off value equal to 0.01. In Figure 5.2 are all 13 possible three-vertex connected directed subgraphs (triads). The triad ID notation in this work is preserving the same notation as on the Figure 5.2 and it is the notation according to [52]. In Figure 5.3 are all 199 possible four-vertex connected directed subgraphs.

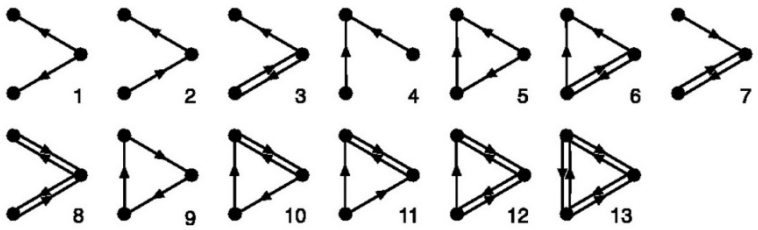


Figure 5.2: All 13 types of three-vertex connected subgraphs.

Now, we will give the mathematical description of the motif in the graph or network G. There are two graphs (networks) H and G with non-empty sets of: vertices, edges and incidence relation. Let H be the real subgraph of $G, H \subset G$. The number of occurrences of graph H in graph G, we define as the frequency of H in G, written like $FH(G)$. Some graph is frequent in G if its frequency in G is higher than cut-off value. Let $\Omega(G)$ be a family of randomized graphs of G (randomized graph has the same number of vertices and same degree sequence [51]). Now we choose n random graphs from $\Omega(G)$ uniformly, $R(G)$. Then we find out the frequency of the certain frequent sub-graph H in G. If the frequency of H in G is higher than its arithmetic mean frequency in n random graphs R_i, where $1 \leq i \leq n$, we call this sample significant and H is network motif for G. Besides the frequency, motifs can be detected by using probabilities. The $p - value$ of the motif is the number of random networks in which a particular motif appeared more frequently than in the original network, divided by the total number of generated random networks. Obviously, the p-value is between 0 and 1. The smaller the p-value of the motif is, the more significant the motif is. Another measure for motif detection is a Z-score . The Z-score for the subgraph H in G can be

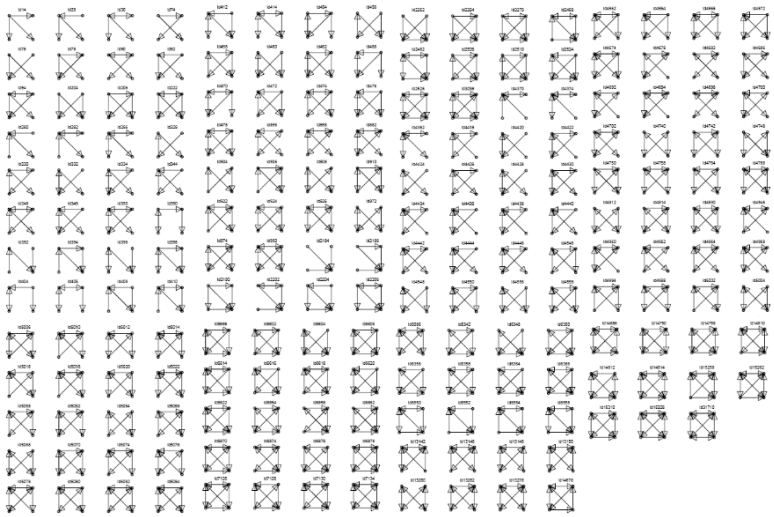


Figure 5.3: Four-vertex connected and directed subgraphs.

calculated using the equation:

$$Z(H) = \frac{F_G(H) - \mu_R(H)}{\sigma_R(H)} \tag{5.1}$$

where $\mu_R(H)$ is the mean and $\sigma_R(H)$ is the standard deviation of frequencies of H in the set of random graphs of G, $R(G)$. The higher the Z-score is, the more significant a detected motif is. Using eq. 1, for each subgraph i, we can calculate the statistical significance which is described as Z-score, Z_i. Furthermore, the SP is the vector of Z-scores normalised to length 1:

$$SP_i = \frac{z_i}{\sqrt{\sum_i z_i^2}}. \tag{5.2}$$

5.4 Experiment

5.4.1 Datasets and Networks Construction

In our study, we examined five literary works. Our dataset contains five different texts; four books: *Mama Leone* (ML), *The Return of Philip Latinowicz* (PL), *The Picture of Dorian Gray*, (DG), *Bones*, (BO) and one forum *Narodne novine* (NN). All the books were written in or have been translated into Croatian. The web forum is selected as a representative of a different text genre in order to verify whether the observed properties are also valid for more relaxed genres besides those strictly for the literature. The datasets are different in the size as well as in the size of the vocabulary (Table 1).

The texts were cleared of the index of contents, the authors' biographies and page numbers. Then we constructed directed co-occurrence networks (word-adjacency networks) in a way that each word represents a vertex, and the two words that follow one another establish the edge.

5.4.2 Network Motifs Analysis

To analyse the motifs in networks we used the FANMOD tool [55]. FANMOD can search for motifs of three to eight vertices sizes using the rand-esu algorithm [20], which is much faster than similar tools, and the advantage is that it has a simple graphical interface and it is very intuitive to

Dataset	Number of words	Number of vertices (N)	Number of edges (K)
ML	86,043	12,416	52,012
PL	28,301	9,166	22,344
DG	75,142	14,120	47,823
BO	199,188	25,020	106,999
NN	146,731	13,036	55,661

Table 5.1: The statistics for the corpus of 10 books.

use. The first step is the preparation of the input data: conversion of words to integers, where every number represents one vertex uniquely in the network, hence two integers in a line form an edge. Every line must contain at least two integers and a maximum of up to five integers. FANMOD provides the possibility to choose whether the networks have directed, undirected or coloured edges or vertices. We used directed uncoloured networks. The algorithm options frame must be adjusted prior to running the algorithm itself. The options' frame includes: the set of the subgraph size and the setting of the switch between full enumeration and enumeration on a few samples. Motifs are identified through the comparison of frequencies in the original network and those in a random network so it is important to determine the number of random networks. It can be set up in the random networks frame in the box named 'Number of networks'. The default value for this is 1,000 networks but it can be increased if necessary. In this frame there are some important parameters: the parameter 'exchanges per edge' (showing how many times the program goes through the edges) should be increased only if our results (output after the first reading) for a random network are very similar to the results for the original network. The parameter 'exchange attempts' - if in the results there appears a small number of successful replacements, then we need to increase it, but it is important to bear in mind that if we have a few successful replacements it may mean that something is wrong with the network. FANMOD produces results in terms of: Z-scores, p-values and frequencies. When we analyse the results, it is desirable to obtain as much as possible undefined Z-scores. If we have a lot of undefined Z-scores, it is not possible to determine which motif is significant (because the greater the Z-score is, the greater significance of this motif is). So if we have a lot of undefined Z-scores we have to increase the number of random networks, which will slow down the algorithm. In the output file format is advisable to include an ASCII - text option for the easier reading of the results, and in HTML format for the presentation of the results. We calculate Z-scores for all triads in all five networks using FANMOD. After that we calculate TSP according to the Eq. 5.2.

5.5 Results

The frequencies of all possible triads for five networks are presented in Figure 5.4. In general, the triad frequencies behave similarly for all five networks. Therefore the Croatian language is comparable with other languages [2, 52]. Still, it is possible to identify differences between data source on ID3#1 and ID3#5 on the linear scale and ID3#9, ID3#11 and ID3#13 on the logarithmic scale.

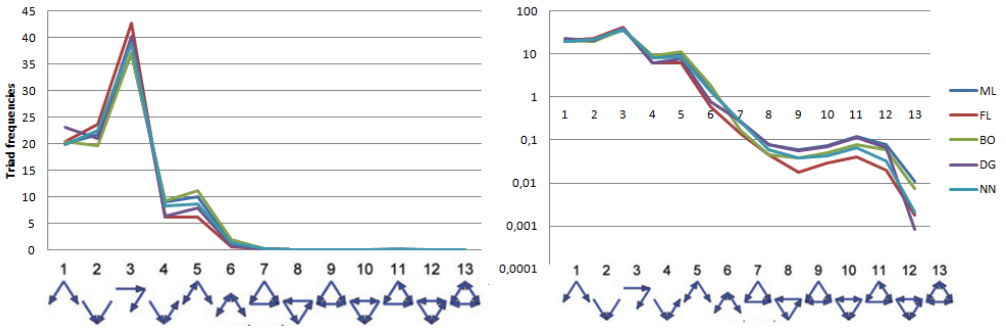


Figure 5.4: The frequencies of the triads for 5 datasets presented on the linear scale (left) and on the logarithmic scale (right).

Furthermore, we analyse TSP in order to detect which triads are significantly overrepresented (motifs) and which triads are significantly underrepresented (anti-motifs) and to compare it across the five different datasets. The results are presented in the diagram shown in the Figure 5.5.

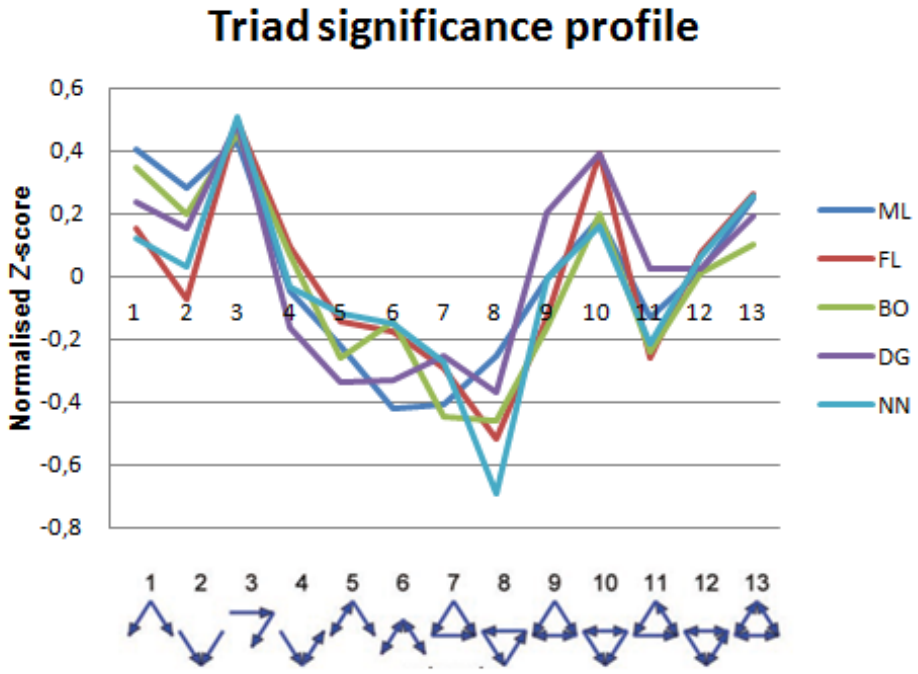


Figure 5.5: Triad significance profile for 5 datasets.

There are several significantly overrepresented triads (ID3#1, ID3#3, ID3#10 and ID3#13). Triads with two edges (ID3#1 and ID3#3) are, based on the other reported results [2, 52], expected to be overrepresented in language networks. However, in our results, triads ID3#10 and ID3#13

are not likely to be overrepresented in language. It seems that this is inherent to languages with a free word-order such as Croatian. For example for three vertices of words: *jako* (much), *ga* (him), *voli* (loves); in Croatian language it is possible to have all six pairs of words (even triplets) as it is shown in Figure 5.6. In opposite, in English language is impossible to have 'him loves' as a part of the sentence.

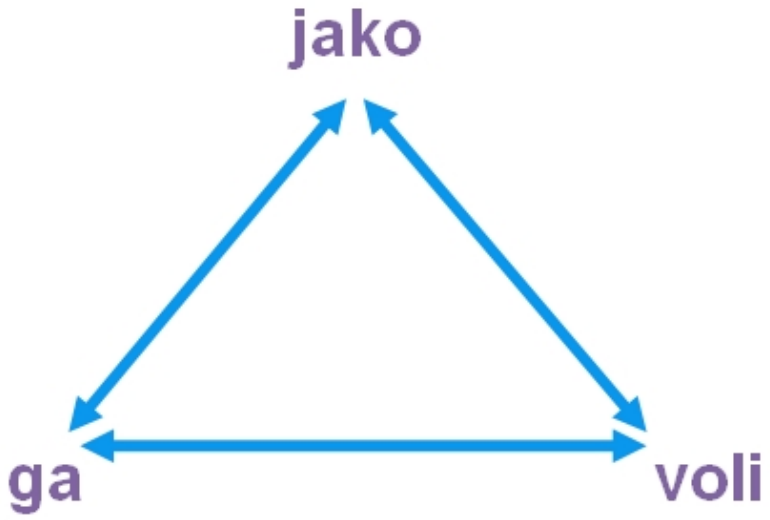


Figure 5.6: An example of the triad with ID3#13 in Croatian language.

5.6 Conclusion

In this Chapter we present the results of the network motifs analysis of Croatian literature. Motifs are used to detect structural similarities between directed networks of four books and one forum. We analyse triad significance profile in five different texts represented as directed co-occurrence networks. The results show that Croatian language networks have similar triad significance profiles with other already analysed languages. Generally, in all language networks triads with two edges are overrepresented, while triads with three edges are underrepresented. For the Croatian language, there is an exception with three-edge triads ID3#10 and ID3#13 which are overrepresented. The overrepresentation of three-edge triads is caused by the free word-order nature of Croatian language. It seems that motif-based analysis of the language networks is sensitive to the word order and syntax rules. And maybe it is possible to use it for the fine-grained differentiation of languages. Therefore, we will perform motif-based analysis of language networks for different languages. We will also include syntax networks and sub-word level networks (syllable networks, grapheme networks) in the analysis. Finally we plan to analyse the presence of the four-vertex motifs in language networks in order to see if they can be interpreted by the semantic relations in the polysemy and synonymy of words.

6. LaNCoA: A Python Toolkit for Language Networks Construction and Analysis

6.1 Abstract

In this Chapter we describe LaNCoA, Language Networks Construction and Analysis toolkit implemented in Python. The toolkit provides various procedures for network construction from the text: on the word-level (co-occurrence networks, syntactic networks, shuffled networks), and on the subword-level (syllable networks, grapheme networks). Furthermore, we implement functions for the language networks analysis on the global and local level. The toolkit is organized in several modules that enable various aspects of language analysis: analysis of global network measures for different co-occurrence window, comparison of networks based on original and shuffled texts, comparison of networks constructed on different language levels, etc. Text manipulation methods, like corpora cleaning, lemmatization and stopwords removal, are also implemented. For the basic network representation we use available NetworkX functions and methods. However, language network analysis is specific and it requires implementation of additional functions and methods. That was the main motivation for this research.

6.2 Introduction

The study of graphs and networks plays an important role in various research domains. The advent of the computer age increased the interest in the large real-world networks that are studied as complex networks. These networks exhibit specific topological properties (high clustering coefficient, small diameter, community structure, one or several giant components, hierarchical structure, heavy tail degree distribution, etc.). Various classes of complex networks have been analyzed, such as for example: technological networks, biological network, information networks or social networks [67]. One possible class includes language networks as well.

Various construction rules may be applied in order to construct a network from the text. The usual way is to construct networks of word co-occurrences [6, 10, 11, 18, 22] or syntactic networks [6, 17, 23, 24, 56, 60, 61]. There are also experiments with shuffled (randomized) networks [4, 11, 15, 17, 19]. Furthermore, syllables networks [31], phoneme networks [26] or semantic networks [3] can be constructed as well. Additionally all these networks can be constructed as

undirected or directed, unweighted or weighted. In most cases the best way to represent the text is to chose directed and weighted variant of the network [18]. There are various software, tools and packages designed and developed for the task of complex networks analysis: Gephi [29], NodeXL [63], SNAP [57], Cytoscape [70], NetworkX package [8] for Python, igraph package [62] for C, Python and R. All these tools enable calculating standard global network measures (such as average clustering coefficient, average shortest path length, diameter, average degree, degree distribution, density, modularity, assortativity, etc.) and the local network measures (different centrality measures: degree, betwenness, closeness, eigenvector, etc.). Some of the tools (for example Gephi, NewtorkX, iGraph, SNAP) provide network analysis on the meso-scale level with implemented algorithms for community detection. Also there are some tools designed and focused only on one aspect of the network analysis, for example GraphCrunch [64] for graphlet analysis, GRAAL [66] for graph and network alignment or FANMOD [54] for motif analysis.

However, there is no software specialized for tasks of language network construction and analysis. Our main motivation was to implement a simple toolkit that provides various language network construction possibilities and suitable network analysis functions. Furthermore this toolkit can be used for various NLP applications, such as for example keyword extraction task [58, 59, 72] or text type classification [65].

The LaNCoA toolkit is focused mainly on the language networks construction task which includes various methods for the corpora manipulation (text preprocessing and cleaning, lemmatization indexlemmatization, shuffling procedures and preparation for the language networks construction) and procedures for generation of various word-level and subword-level networks directly from the given corpora. The toolkit also enables complex network analysis in terms of calculating all important global and local network measures , network and text content analysis, and various plotting data possibilities. To some extent, the LaNCoA toolkit uses existing functions from the NetworkX Python package as a basic foundation for some more specific network construction and analysis tasks. Furthermore, there are some measures important for weighted and directed networks that are not implemented in the standard network-manipulation packages which are therefore implemented in the LaNCoA toolkit, such as the selectivity measure, network reciprocity , network entropy , inverse participation ratio , and link overlap measures.

The Chapter 6 is structured as follows. In Section 6.3 we describe the language networks. In Section 6.4 we give a short overview of the complex networks analysis task. Then we present the LaNCoA toolkit in Section 6.5 and we describe LaNCoA toolkit applications in Section 6.6. We give a conclusion remarks in Section 6.7.

6.3 Complex Network Analysis Task

A complex network is modeled as a graph G. A graph $G = (V, E)$ consists of a collection of vertices, or vertex set, V and a collection of edges, or edge set, E. In the complex network approach vertices are called nodes and edges are called links. The study of networks can be classified in three levels: global (macro-scale) level, meso-scale level and local (micro-scale) level.

The study at the macro level attempts to understand the global structure of a network. At this level, relevant parameters are average degree, degree distribution, average path length, average clustering coefficient, density, modularity, assortativity, etc. An the meso-scale level the interaction between nodes at short distances are studied. This includes community detection or analysis of small subgraphs such as motifs or graphlets . At the micro level the study is focused on the behavior of single nodes. Identification of the important nodes in the network using different centrality measures or just deterring degree, strength, clustering coefficient or betweenness and other parameters of a single node . In [67] a detailed overview of all network measures and formulas is given.

6.4 Language Networks

Written, as well as spoken language can be modeled via complex networks where the lingual units (e.g. words) are represented by vertices and their linguistic interactions by links. Language networks are a powerful formalism to the quantitative study of language structure at various language sublevels . Complex network analysis provides mechanisms that can reveal new patterns in complex structure and can thus be applied to the study of patterns that occur in the natural languages. Thus, complex network analysis may contribute to a better understanding of the organization, structure and evolution of a language.

On the word-level, text can be represented as a complex network of linked words: each individual word is a node and interactions amongst words are links [6]. The interactions can be derived at different levels: structure, semantics, dependencies, etc. On the subword-level, syllable or grapheme networks can be constructed, where nodes can be represented by syllables or graphemes, while their dependencies (e.g. positions of syllables within words or graphemes within syllables) are links [31].

The properties of the co-occurrence networks are derived from the word order in texts [6, 10, 11, 18, 22]. Commonly they rise from the simple criterion such as co-occurrence of two words within a sentence, or text; or as co-occurrence of words within the given co-occurrence window. In the networks where the linkage is limited to the sentence borders during the construction, the sentence boundary can be considered as the window boundary too.

The syntactic networks are constructed using syntactic dependencies relations. Syntactic dependencies between words are formally expressed by dependency grammar (e.g. set of productions (rules) in a form of a grammar). The dependency grammar is used to present the syntactic relationships from sentence in a form of syntactic dependency tree. The properties of the syntactic networks are analyzed in [6, 17, 23, 24, 56, 60, 61]. The results suggests that modelling human language using syntactic networks is important for language analysis because not all of the properties of the text structure are captured within co-occurrence networks.

For the purpose of better understanding of the language structure, one approach to address the questions of the word order in the language is to compare networks constructed from normal texts with the networks from randomized or shuffled texts [4, 11, 15, 17, 19]. Networks constructed from such shuffled texts are commonly regarded as shuffled networks .

Syllable and grapheme networks are important for studying structure of a language at the subword-level. In the syllable networks , nodes are represented by syllables and a link between two syllables can be established if they belong to the same word or if they are neighbors in the word. [31]. The same principle applies to the grapheme networks , where two graphemes are linked if they co-occurre as neighbors within a word or a syllable.

Figure 1 present different construction rules for stated language network types.

6.5 The LaNCoA Toolkit Overview

LaNCoA is free and open source software licensed under the General Public License version 2. The source code of a working version is available for download from the official GitHub repository at https://github.com/domargan/LaNCoA.

All of the LaNCoA functionalities work for any of the languages written in any set of graphemes based on the letters of the classical Latin alphabet (Latin script). Only Latin script languages are supported (commonly used by about 70% of the world's population).

The LaNCoA toolkit is implemented in Python programming language. Python is an excellent tool for scanning and manipulating textual data and also provides various packages and libraries for scientific computations and data visualization. One of those is the popular NetworkX package, designed for exploration and analysis of complex networks and network algorithms. Our goal was

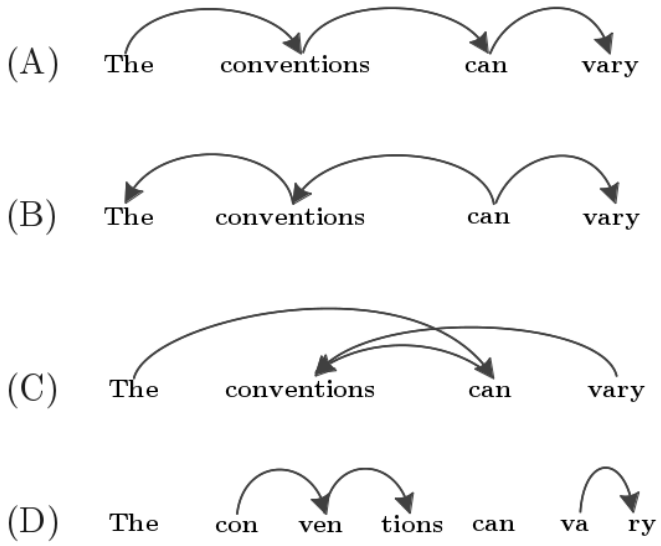


Figure 6.1: Language networks construction rules presented on one toy-example sentence "The conventions can vary.": (A) Co-occurrence network, (B) Syntactic dependency network, (C) Shuffled network, (D) Syllable network.

to utilize basic NetworkX functions to develop extra procedures suitable for language network construction and analysis. We have also implemented functions for calculation of some non-standard general complex network measures. We have based our plotting procedures on proven quality matplotlib library for producing visualization figures.

Toolkit is divided into six modules that enable various aspects of language and text corpora analysis: i) corpora manipulation, ii) language networks generation, iii) single language network analysis, iv) multiplex language networks analysis, v) content analysis, and vi) data plotting. Modules are grouped into two main parts: network construction and network analysis. Modules provide procedures for tasks such as corpora cleaning, utilization of different network construction principles, analysis of global and local network properties, comparison of networks based on original and shuffled corpora, comparison of networks constructed on different language levels, etc. The generalized architectural structure of our toolkit is visualized and presented in Figure 6.2. In this Section we describe each module's function individually.

6.5.1 Network Construction

Network construction part of our toolkit consists of two modules: the corpora manipulation module and the network generation module.

Corpora Manipulation Module

Corpora manipulation module can be used for several tasks with focus on various functions used to manipulate textual corpora. All of the tasks are optional and can be performed independently by the user's choice. Implemented LaNCoA functions are the following:

Corpus cleaning and Unicode normalization

It is important to have clean and high quality corpus before the process of the network generation, so the user can utilize LaNCoA to clean the corpus from the unwanted characters or data . LaNCoA also supports the usage of unclear and "dirty" textual data, but the noise-cleaning of the corpus is recommended, since it can greatly reduce the risks of badly constructed networks or network pollution. All UTF-8 characters which are not defined as letters or numbers of the classical Latin

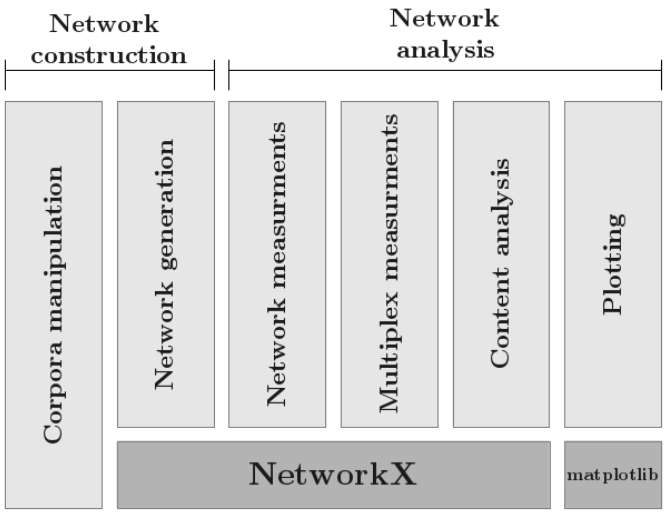


Figure 6.2: LaNCoA architecture.

alphabet can be removed from the textual corpus . On the other hand, any set of those UTF-8 characters (or all of them) can be kept (preserved) in the text by users' choice. In addition, the NFKD unicode normalization [73, 74] of all Latin script letters can optionally be performed directly in the process of corpus cleaning. Compatibility decomposition replaces the code points of a base letter into a single precomposed letter. For example, unicode letters 'ć' or 'š' can be normalized into 'c' and 's' characters.

Removal of stopwords from a corpus

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,...). Examples of stopwords are: 'is', 'but', 'and', 'which', 'on', 'any', 'some'. Stopwords from any language based on the Latin script can be removed by providing adequate text file containing the list of stopwords .

Lemmatization of a corpus

Lemmatization is the process by which single words are reconducted to their citational form. For instance the word 'networks' is converted into its standard form 'network'. Lemmatization, along with the morphological analysis, is the foundation of all the processes involved in language normalization. Lemmatization can be performed for any language based on the Latin script by providing adequate text file containing the list of all word form-lemma pairs, since the lemmatization in LaNCoA is based on the find-and-replace principle .

Text shuffling

Co-occurrence complex networks properties are derived from the word order in texts. Commonly, the shuffling procedure randomizes the words in the text, transforming the text into the meaningless form. Shuffling procedures destroy the sentence and text organization in a way that the standard word-order and syntax of the text is eradicated. As expected, the typical word collocations and phrases are completely lost, as well as the forms of the morphological structures and local structures of words' neighborhood. We implemented two different shuffling principles: shuffling on the sentence level and shuffling on the whole text level. The vocabulary size, word and sentence frequency distributions stay preserved in both shuffling procedures. Additionally, in the sentence level shuffling approach, the sentence structure of the text and the number of words per sentence are also preserved . In the text-level shuffling, the original text is randomized by shuffling the words

and punctuation marks over the whole text. This approach also changes the number of words per sentence.

Network Generation Module

LaNCoA's network generation module can be used for the generation of complex language networks directly from corpora or from other language networks. It can be used for several independent tasks of building networks on word and subword-level Networks can be generated as weighted or unweighted, as well as directed or undirected . All generated networks can be saved for later use in the standard edgelist file format . Implemented functions are the following:

Co-occurrence networks generation

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. Words are also linked according to the optional usage of specified delimiters (e.g. punctuation marks). In the networks where the linkage is limited to the sentence borders during the construction, the sentence boundary is then the window boundary too. In the networks without delimiters, words are linked within a given co-occurrence window regardless of being in different sentences. Standard approach is to limit the co-occurrence window size within the sentence delimiters, but a user may or may not specify any type of delimiters (any UTF-8 character). 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. Co-occurrence networks can be generated directly from the raw text data that does not necessary conform to rules of grammar or orthography. 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 Figure 3.

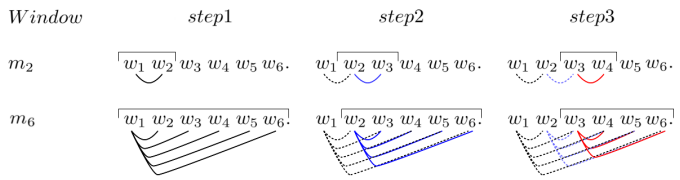


Figure 6.3: An illustration of 3 steps in a network construction with a co-occurrence window m_n of sizes $n = 2$, and $n = 6$. $w_1 ... w_6$ are words within a sentence.

Syntactic networks generation

The syntactic structure of language is captured through syntactic dependency relations between pair of words in a sentence : the head word – the governor of relationship and the dependent word - the modifier. Syntactic dependencies between words are formally expressed by dependency grammar which is used to represent the syntactic relationships from sentence in a form of syntactic dependency tree. The sentences boundaries are preserved, since the syntactic dependency is inherent to the sentence. The weight of the link between two nodes is proportional to the overall frequencies of the corresponding words within a syntactic dependency tree. User must provide a corpus in a form of syntactic dependency treebank file written in the CoNLL-X format.

Syllable and grapheme networks generation

The syllable networks are constructed from the co-occurrence of syllables within words. Syllable list is obtained from the dictionary file already containing syllabified words . The weight of the link between two syllables is proportional to the overall frequencies of the corresponding syllables co-occurring within words from a text. Syllable networks can be generated directly from the raw text corpora. The structure of grapheme networks depends on a existing network of syllables. Two

graphemes are linked if they co-occur as neighbours within a syllable. The weight of the link between two graphemes is proportional to the overall frequencies of the corresponding graphemes co-occurring within syllables from a syllable network .

Word-list subnetwork and word-ego subnetwork generation

Two types of word-level subnetworks can be generated from existing co-occurrence or syntactic networks: word-list network and word-ego network . Word-list network is a simple subnetwork based on a provided list of words. Specified nodes (words) and corresponding links between them are extracted from the original network. Word-ego network is a subnetwork of neighbors centered at one specified node (word) within a given radius. These subword-level networks can be generated to examine the networks of semantic importance, word's predecessors, successors, or entire word neighborhood of keywords within a given radius.

6.5.2 Network Analysis

Network analysis part of the LaNCoA toolkit consists of four modules: single network analysis module, multiplex network analysis module, content analysis module and data plotting module.

Single network analysis module

Single network properties can be analyzed on global and local scale. LaNCoA uses some calculation methods implemented in the NetworkX Python package. These include standard basic network features, for e.g., the average path length, diameter and radius, global and local clustering coefficient, network transitivity, and network density. In addition to NetworkX's procedures used for calculation of classic network properties, LaNCoA provides several other procedures for calculation of non-standard network measures, such as selectivity and inverse participation ratio (both available for directed and undirected networks), calculation of network entropy based on the degree, strength, selectivity and inverse participation distributions, and network reciprocity.

Multiplex network analysis module

LaNCoA provides some simple functions for analysis of multiplex networks . Multiplex network is network in layers, and with connections between layers; the interconnections between layers are only between a node and its counterpart in the other layer (the same node) . This module enables overlap analysis of two different separated networks consisted of the same sets of nodes. Implemented functions, for example, enable calculation of the Jaccard distance of two different networks, as well as the total and total weighted link overlap measures .

Content analysis module

Content analysis implies the examination of the text corpora's content (e.g. role of the individual words) by using the complex network environment. This module provides several ways of text analysis by calculating simple network statistics, such as the top n words, syllables or graphemes with the largest number of different individual neighbors, calculation of the most frequent word-pair relations, the distance between given words in the network environment, or calculation of the centrality measures for all of the words within a corpus.

Data plotting module

LaNCoA provides functions for plotting of the network's data by utilizing the methods from the matplotlib Python library. Users can generate various 2D figures based on the calculated network measures. Such figures include rank plots for the directed (in- and out-) or undirected degree, strength and selectivity distribution values of multiple networks on the same scale, as well as the degree, strength and selectivity histograms and scatter plots. It is also possible to generate the plots describing the dynamic growth of a network regarding the number of connected components, presenting the ratio of newly 'read' unique words (or syllables or graphemes) and

the corresponding number of components in a given point of time in the process of co-occurrence network construction.

6.6 The LaNCoA Toolkit Applications

We have used the LaNCoA toolkit in several of our experiments where we have worked with the language networks.

In [18] we presented the results of our first experiment with the Croatian co-occurrence language networks. In this experiment we constructed 30 different co-occurrence networks, weighted and directed, from the corpus of literature, containing 10 books written in or translated into the Croatian language. We examined the change of network structure properties by systematically varying the co-occurrence window sizes, the corpus sizes and removing stopwords. We used the LaNCoA toolkit for all these network construction tasks.

In [14] we compared Croatian, English and Italian language networks based on the same five books. We performed lemmatized and non-lemmatized network construction with and without stopwords using the LaNCoA toolkit.

In another experiment [21] we addressed the problem of Croatian text complexity by constructing the linguistic co-occurrence networks form normal texts and shuffled text. In this experiment we have tested whether complex network measures can differentiate between normal and shuffled texts. We employed various methods from the LaNCoA toolkit for calculating the network measures and generating various plots in order to find the differences between two classes of networks. In [50] we extended this research by introducing additional shuffling procedure: the sentence-level shuffling procedure and by introducing a node selectivity as a new complex network measure. All shuffling procedures and network construction tasks were performed with the LaNCoA toolkit.

Furthermore, we used the LaNCoA toolkit for various experiments with the selectivity measure. In [71] we compared language networks from Croatian literature and blogs. In [58, 59, 72] we analysed the potential of the selectivity measure for the keyword extraction task. We also used the LaNCoA toolkit for the Croatian language networks construction for the purposes of the network motif analysis of Croatian literature performed in [69]. In [65] we used methods from the LaNCoA toolkit to generate 150 different weighted and directed networks and to calculate local and global network measures used in the task of text classification.

6.7 Conclusion

In this Chapter we presented an overview of the LaNCoA toolkit for language networks construction and analysis. Currently, its basic functionalities rely on the corpora manipulation and language network construction methods implemented in the two separate modules. Another set of modules provide methods for the network analysis task. These modules employ some of the basic methods that already exists in the NetworkX package. However there is a set of functions for the network analysis not covered by the standard network-manipulation packages. Among them are certain functions that deals with the measures for the weighted and directed networks. These functions are of special interest for the language networks analysis and we implemented them in our toolkit.

The LaNCoA toolkit is in the early stage of development and there is still place for major improvements, especially in the network analysis tasks suited for the language networks. However, we managed to use this toolkit successfully in all of the language network experiments that we performed. For the future work, we plan to implement simple and robust user interface. In addition, we would like to develop some more specific language-oriented network analysis functions and also make improvements in the existing code whenever it is possible.

II Applications

7. An Overview of Graph-Based Keyword Extraction Methods and Approaches

7.1 Abstract

The Chapter surveys methods and approaches for the task of keyword extraction. The systematic review of methods was gathered which resulted in a comprehensive review of existing approaches. Work related to keyword extraction is elaborated for supervised and unsupervised methods, with a special emphasis on graph-based methods. Various graph-based methods are analyzed and compared. The Chapter provides guidelines for future research plans and encourages the development of new graph-based approaches for keyword extraction.

7.2 Introduction

Keyword extraction (KE) is tasked with the automatic identification of a set of the terms that best describe the subject of a document [75, 79, 80, 87, 96, 99, 106, 110, 122, 132]. Different terminology for defining the terms that represent the most relevant information contained in the document is used: key phrases, key segments, key terms or just keywords . All listed variants have the same function - to characterize the topics discussed in a document [110]. Extracting a small set of units, composed of one or more terms, from a single document is an important problem in Text Mining (TM), Information Retrieval (IR) and Natural Language Processing (NLP).

Keywords are widely used to enable queries within IR systems as they are easy to define, revise, remember, and share. Keywords are independent of any corpus and can be applied across multiple corpora and IR systems [79]. Keywords have also been applied to improve the functionality of IR systems [79, 84]. In other words, relevant extracted keywords can be used to build an automatic index for a document collection or alternatively they can be used for document representation in categorization or classification tasks [99, 110]. An extractive summary of the document is also the task of many IR and NLP applications and includes automatic indexing, automatic summarization, document management, high-level semantic description, text, document or website categorization or clustering, cross-category retrieval, constructing domain-specific dictionaries, name entity recognition, topic detection, tracking, etc. [79, 92, 103, 111, 125].

While assigning keywords to documents manually is a very costly, time consuming and tedious task, in addition to which, the number of digitally available documents is growing, automatic keyword extraction has attracted the interest of researchers over the last years. Although the keyword extraction applications usually work on single documents, keyword extraction is also used for a more complex task (i.e. keyword extraction for the whole collection [126], the entire web site or for automatic web summarization [131]) . With the appearance of big-data, constructing an effective model for text representation becomes even more urgent and demanding at the same time [100]. State-of-the-art techniques for KE encounter scalability and sparsity problems. In order to circumvent these limitations, new solutions are constantly being proposed. This work presents a comprehensive overview of the common techniques and methods with the emphasis on new graph-based methods, especially regarding keyword extraction for the Croatian language. We systematize the existing state-of-the-art keyword extraction methods and approaches as well as new graph-based methods that are based on the foundations of graph theory. Additionally, the paper explores the advantages of graph-based methods over traditional supervised methods.

The Chapter is organized as follows: firstly, we systematize keyword extraction methods; secondly, we present a brief overview of various measures for network (graph) analysis; thirdly, we describe related work for supervised and unsupervised methods, with special emphasis on graph-based keyword extraction; fourthly, we compare graph-based measures of experiments extracting keywords from Croatian News articles; and finally, we conclude with some remarks regarding network-enabled extraction and turn to brief guidelines for future research.

7.3 Systematization of Methods

Keyword assignment methods can be divided roughly into two categories: (1) keyword assignment and (2) keyword extraction [86, 102, 121, 125] as presented in Figure 7.1 . Both revolve around the same problem - selecting the best keyword. In keyword assignment, keywords are chosen from a controlled vocabulary of terms or predefined taxonomy, and documents are categorized into classes according to their content. Keyword extraction enriches a document with keywords that are explicitly mentioned in text [107]. Words that occurred in the document are analyzed in order to identify the most representative ones, usually exploring the source properties (i.e. frequency, length) [129]. Commonly, keyword extraction does not use a predefined thesaurus to determine the keywords .

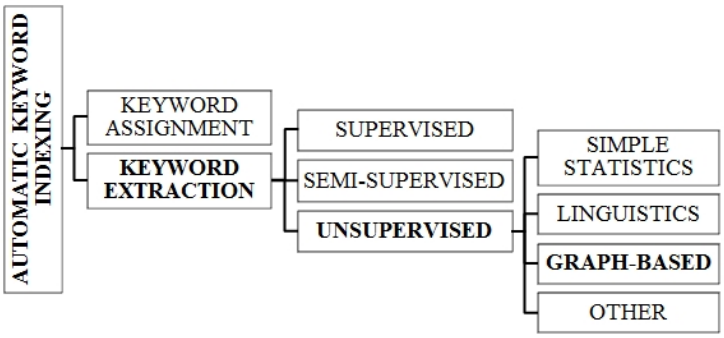


Figure 7.1: Classification of keyword extraction methods.

The scope of this work is calibrated only on keyword extraction methods. Existing methods for automatic keyword extraction can be according to Ping-I and Shi-Jen [82] divided roughly into:

- Statistical Approaches and

- Machine Learning Approaches,

or slightly more detailed in the four categories according to Zhang et al. [129]:

- Simple Statistical Approaches,
- Linguistic Approaches,
- Machine Learning Approaches and
- Other Approaches.

Simple Statistical Approaches comprise of simple methods which do not require the training data. In addition, these methods are language and domain-independent. Commonly, the statistics of the words from a document can be used to identify keywords: n-gram statistics, word frequency, TF-IDF (term frequency-inverse document frequency model, word co-occurrences, PAT Tree (Patricia Tree; a suffix tree or position tree), etc. The disadvantage is that in some professional texts, such as from the health and medical domain, the most important keyword may appear only once in the article (e.g. diagnosis). The use of statistically empowered models may inadvertently filter out these words [82].

Linguistic Approaches use the linguistic properties of the words, sentences and documents. Lexical, syntactic, semantic and discourse analysis are some of the most commonly examined properties, although they are demanding and complex NLP problems.

Machine Learning Approaches consider supervised or unsupervised learning from the examples, but related work on keyword extraction prefers the supervised approach. Supervised machine learning approaches induce a model which is trained on a set of keywords. They require manual annotations of the learning dataset which is extremely tedious and inconsistent (sometimes requesting predefined taxonomy). Unfortunately, authors usually assign keywords to their documents only when they are compelled to do so. The model can be induced using one of the machine learning algorithms: Naïve Bayes, SVM (Support Vector Machines), C4.5, etc. Thus, supervised methods require training data, and are often dependent on the domain. A system needs to re-learn and establish the model every time when a domain changes [94, 115]. Model induction itself can also be demanding and time consuming on massive datasets .

Other Approaches for keyword extraction in general combine all the methods mentioned above. Additionally, sometimes for fusion they incorporate heuristic knowledge, such as the position, the length, the layout features of the terms, html and similar tags, the text formatting information etc.

Vector space model(VSM) is well-known and is the most used model for text representation in text mining approaches [78, 86, 90] . Specifically, the documents represented in the form of feature vectors are located in a multidimensional Euclidean space . This model is suitable for capturing simple word frequency, however structural and semantic information are usually disregarded. Due to its simplicity VSM has several disadvantages [116]:

- the meaning of a text and structure cannot be expressed explicitly,
- each word is independent from other, word appearance sequences or other relations are disregarded,
- if two documents have a similar meaning expressed with different words, similarity cannot be computed easily.

Graph-based text representation efficiently addresses these problems [116]. A graph is a mathematical model, which enables the exploration of the relationships and structural information very effectively . More about the graph representations of text is discussed in Section 3, and in [22, 59, 105, 116, 124]. For now, in short, document is modelled as graph where terms (words) are represented by vertices (nodes) and their relations are represented by edges (links). The taxonomy of the graph-enabled keyword extraction methods is presented in Figure 7.4.

The edge relation between words can be established on many principles exploiting different scopes of the text or relations among words for the graph's construction [105, 116]:

- co-occurrence relations - connecting neighboring words co-occurring within the window of a fixed size in text; or connecting all words co-occurring together in a sentence, paragraph, section or document (adding them to the graph as a clique[1]);
- syntax relations - connecting words according to their relations in the syntax dependency graph;
- semantic relations - connecting words that have similar meanings, words spelled the same way but have different meanings, synonyms, antonyms, homonyms, etc;
- other possible relations - for example, intersecting words from a sentence, paragraph, section or document, etc.

There are various possibilities for the analysis of a network structure (topology) and we will focus on the most common - network structure of the linguistic elements themselves using various relations: semantic, pragmatic, syntax, morphology, phonetic and phonology. More precisely, in this work we narrow the scope of the study to (1) **co-occurrence** [3], (2) **syntactic** [17], (3) **semantic** [124] and (4) **similarity networks** [105].

7.3.1 Graph Types

The formal definition of a graph according to graph theory is given in Section 3. Here we broadly discuss the classification of a graph-based method which can be established on the (1) **vertices** or (2) **edges** [105].

In vertex representation models, vertices represent advanced concepts which can be **atomic** (one component; also called homogenous) or **multiple** (more than two components; also called heterogeneous). The homogeneous graph model is usually used for the representation of grammatical associations between words or semantic similarities [81, 98]. Additionally, vertices can also be weighted or unweighted which conditions the representation model, which is respectively (1) **weighted** or (2) **unweighted** graph. Weighted vertices in this case commonly indicate the importance of the vertex in the graph, and different measures (explained in Section 3) are used to calculate the importance of the vertex. The measures and algorithms listed in Table 1, very often take into account the number of edges, the weight of vertices which are connected by the edge, etc.

Between two vertices, relationships can be established by edges. In edge representation models (graphs) graphs can be either (1) **directed** (called digraph, e.g. for word order in text) or (2) **undirected** (for connecting related words). Edges can also be (1) **weighted** or (2) **unweighted**, depending on relationships between vertices. In a language complex network, weight could be the distance of two words in paragraphs or text or the frequency of word pairs' co-occurrence. Beside weights, edge models can be (1) **labeled** or (2) **unlabeled**. It is almost conventional to explain the relationships or rules between related vertices by the edge label in many graph models in computer science (e.g. Entity-Relationship). In related work of graphs in the language's edge label can denote POS (part of speech), grammatical rule of word, etc.

There are also more complex models that are represented by combinations of the previously described models or parts of their structure. These are: (1) **multigraphs** - this model allows a connection with a plurality of different edges, and also a vertex connection with itself, (2) **hipergraph** - one connection can be established with any number of vertices; edges are not binary relations, (3) **multiplex** - a multilayer graph which shares the same vertices at all levels, and has edges between levels that are achieved by connecting only the same vertices.

An example of such a model is the multiplex of many realizations of the same text, always containing the same set of words interlinked with different edges: as direct neighbors, co-occurrence in the sentence, syntax dependencies, etc.

The classifications of graph types with all previous described features are shown in Figure 7.2 according to concepts, weight, direction or label for a vertex or edge representation model. The

[1] Clique is a subgraph of a graph in which every two vertices are connected (a subgraph which is a complete graph).

classifications of advanced graph models are shown in Figure 7.3.

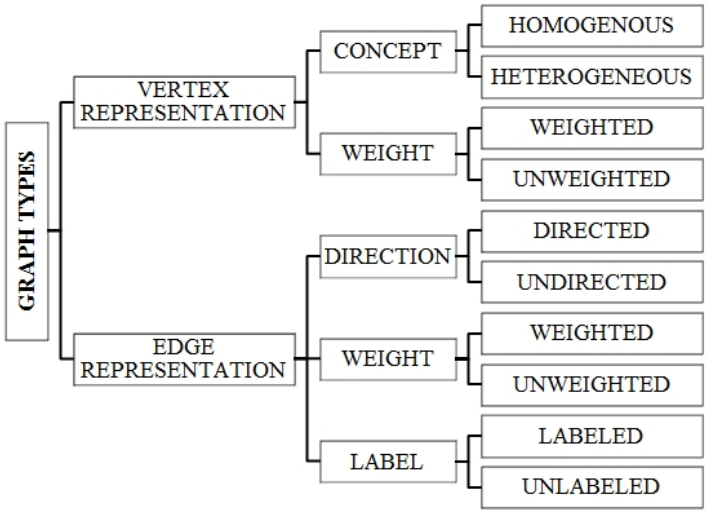


Figure 7.2: Classification of graph types.

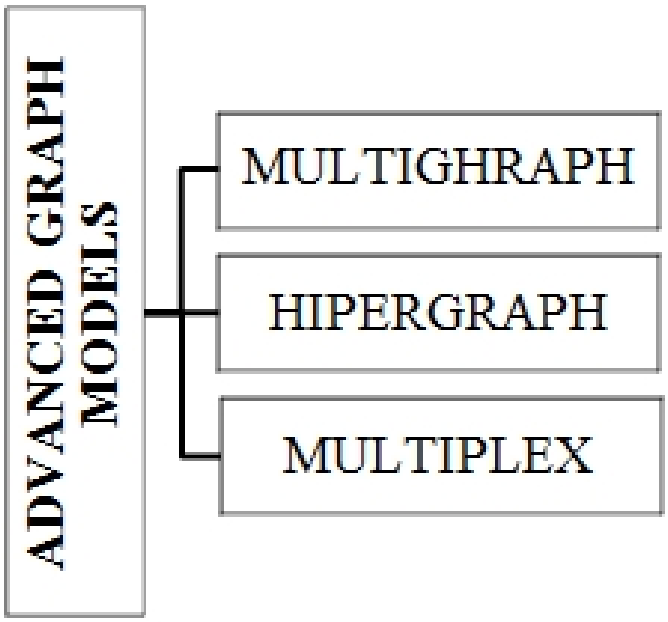


Figure 7.3: Classification of advanced graph models.

7.4 Graph-based Centrality Measures

This Section defines some basic concepts from graph theory and the centrality measures necessary for understanding the graph-based approach. More details about graph measures can be found in [3, 19, 67, 83].

A graph is an ordered pair $G = (V, E)$ where V is the set of vertices and $E \subseteq V \times V$ is the set of edges. A graph is directed if the edges have a direction associated with them. A graph is weighted if there is a weight function ω that assigns value (real number) to each edge. We use $N = |V|$ and $K = |E|$ as shorthand for the number of vertices and edges in a graph.

A path in a graph is a sequence of edges which connects a sequence of vertices which are all distinct from one another. A shortest path between two vertices u and v is a path with the shortest length and it is called the distance between u and v.

In the graph theory centrality measures refer to indicators which identify the most important vertices within a graph and that approach is used for the task of ranking the vertices. In the domain of keyword extraction various centrality measures are proposed and used for the task of ranking the words in a text.

Centrality measures are local graph measures, focused on a single vertex and its neighborhood. The neighborhood of a vertex v in graph G is defined as a set of neighbors of a vertex v and is denoted by $N(v)$. The neighborhood size is the number of immediate neighbors to a vertex. The number of edges between all neighbors of a vertex is denoted by $E(v)$. In the directed graph, the set of $N_{in}(v)$ is the set of vertices that point to a vertex v (predecessors) and set of $N_{out}(v)$ is the set of vertices that vertex v points to (successors) .

The clustering coefficient of a vertex measures the density of edges among the immediate neighbors of a vertex . It determines the probability of the presence of an edge between any two neighbors of a vertex. It is calculated as a ratio between the number of edges E_i that actually exist among these and the total possible number of edges among neighbors:

$$c(v) = \frac{2E(v)}{|N(v)|(|N(v)|-1)}. \tag{7.1}$$

The degree $d(v)$ of a vertex v is the number of edges at v; it is equal to the number of neighbours of v.

In a directed graph, the in-degree of a vertex v, $d^{in}(v)$ is defined as the number of inward edges from a vertex v. Analogously, the out-degree of a vertex v, $d^{out}(v)$ is defined as the number of outward edges from a vertex v.

The degree centrality $C_d(v)$ of a vertex v is defined as the degree of the vertex. It can be normalized by dividing it by the maximum possible degree $N-1$:

$$C_d(v) = \frac{d(v)}{N-1}. \tag{7.2}$$

In the directed graph the in-degree centrality of the vertex v is defined as in-degree of the vertex (normalized by dividing it by the maximum possible degree $N-1$) :

$$C_d^{in}(v) = \frac{d^{in}(v)}{N-1}. \tag{7.3}$$

The out-degree centrality $C_d^{out}(v)$ of a vertex v is defined analogously .

The strength of the vertex v is a sum of the weights of all the edges incident with the vertex v :

$$s(v) = \sum_u w_{vu}. \tag{7.4}$$

In the directed network, the in-strength $s^{in}(v)$ of the vertex v is defined as the sum of all weights of inward edges from a vertex v:

$$s^{in}(v) = \sum_u w_{uv}. \tag{7.5}$$

The out-strength $s^{out}(v)$ of a vertex v is defined analogously .

The selectivity measure is introduced in [19] . It is an average strength of a vertex. For the vertex v the selectivity is calculated as a fraction of the vertex strength and vertex degree:

$$e(v) = \frac{s(v)}{d(v)}. \tag{7.6}$$

In the directed network, the in-selectivity of the vertex v is defined as :

$$e^{in}(v) = \frac{s^{in}(v)}{d^{in}(v)}. \tag{7.7}$$

The out-selectivity $e^{out}(v)$ of a vertex v is defined analogously .

The closeness centrality $C_c(v)$ of a vertex v is defined as the inverse of farness, i.e. the sum of the shortest distances between a vertex and all the other vertices in a graph. Let d_{vu} be the shortest path between vertices u and v. The normalized closeness centrality of a vertex v is given by:

$$C_c(v) = \frac{N-1}{\sum_{v \neq u} d_{vu}}. \tag{7.8}$$

The betweenness centrality $C_b(v)$ of a vertex v quantifies the number of times a vertex acts as a bridge along the shortest path between two other vertices . Let σ_{ut} be the number of the shortest paths from vertex u to vertex t and let $\sigma_{ut}(v)$ be the number of those paths that pass through the vertex v. The normalized betweenness centrality of a vertex v should be divided by the number of all possible edges in the graph and is given by:

$$C_b(v) = \frac{2\sum_{v \neq u, u \neq t} \frac{\sigma_{ut}(v)}{\sigma_{ut}}}{(N-1)(N-2)}. \tag{7.9}$$

The eigenvector centrality $C_{EV}(v)$ measures the centrality of a vertex v as a function of the centralities of its neighbors. For the vertex v and constant λ it is defined :

$$C_{EV}(v) = \frac{1}{\lambda} \sum_{u \in N(v)} C_{EV}(u). \tag{7.10}$$

In the case of weighted networks, the equation can be generalized. Let w_{uv} be the weight of edge between vertices u and v and λ a constant. The eigenvector centrality of a vertex v is given by:

$$C_{EV}(v) = \frac{1}{\lambda} \sum_{u \in N(v)} w_{uv} \times C_E(u). \tag{7.11}$$

There are various centrality measures based on the idea of eigenvector centrality defined. The HITS method defines authority $x(v)$ and a hub score $y(v)$ for vertex v. Let e_{vu} represent the directed edge from vertex v to vertex u. Given that each vertex has been assigned an initial authority score $x(v)^{(0)}$ and hub score $y(v)^{(0)}$ as described in [97], HITS iteratively refines these scores by computing:

$$x(v)^i = \sum_{u: e_{uv} \in E} y(u)^{i-1} \tag{7.12}$$

$$y(v)^i = \sum_{u:e_{vu} \in E} x(u)^i \tag{7.13}$$

for $k = 1, 2, ...$

The TextRank centrality is based on the eigenvector centrality measure and implements the concept of 'voting' . The TextRank score of a vertex v is initialized to a default value and computed iteratively until convergence using the following equation:

$$C_{PageRank}(v) = (1-d) + d \sum_{u \in N_{in}(v)} \frac{C_{PageRank}(u)}{|N_{out}(u)|} \tag{7.14}$$

where d is the dumping factor set between 0 and 1 (usually set to 0.85). The TextRank is a modification of a PageRank defined for weighted graphs and used for ranking words in the texts . The equation is:

$$C_{TextRank}(v) = (1-d) + d \sum_{u \in N_{in}(v)} \frac{w_{uv} \times C_{TextRank}(u)}{\sum_{t \in N_{out}(u)} w_{ut}}. \tag{7.15}$$

7.5 Related Work on Keyword Extraction

Although the keyword extraction methods can be divided as (1) **document-oriented** and (2) **collection-oriented**, we are most interested in some of the other systematization in order to get a broad overview of the field. The approaches for keyword extraction can be roughly categorized into either (1) **unsupervised** or (2) unsupervised . Supervised approaches require an annotated data source, while the unsupervised require no annotations in advance. The massive use of social networks and Web 2.0 tools has caused turbulence in the development of new methods for keyword extraction. In order to improve the performance of methods on massive quantities of data (3) **semi-supervised** methods have come into research focus. Figure 7.1 shows the different techniques that are combined into supervised, unsupervised or both approaches.

Two critical issues of supervised approaches are demands to prepare the training data with manually annotated keywords and the bias towards the domain on which they are trained. For this reason in this work, the focus has been shifted towards more unsupervised methods, specifically graph-based methods which have been developed using only the statistics of the source which is reflected into the structure of the graph (network).

7.5.1 Supervised

The main idea of supervised methods is to transform keywords extraction into a binary classification task - word is either a keyword or not. Two typical and well-known systems for supervised keyword extraction, which set the boundaries of the research field are Kea (Witten et al., 1999 [125]) and GenEx (Turney, 1999 [121, 125]). The most important features for classifying a keyword candidate in these systems are the frequency and location of the term in the document. In short, GenEx uses Quinlan's C4.5 decision tree induction algorithm to his learning task, while Kea uses Naïve Bayes algorithm for training and keyphrase extraction. GenEx and Kea are extremely important systems because, in this field of keyword extraction, they set up the foundation for all other methods that were developed after, and have become the state-of-the-art benchmark for evaluating the performance of other methods.

Hulth (2003) in [92] explores the incorporation of linguistic knowledge into the extraction procedure and uses Noun Phrase chunks (NP) (rather than term frequency and n-grams), and adds the POS (Part-of-Speech) tag(s) assigned to the term as a feature. In more details, extracting

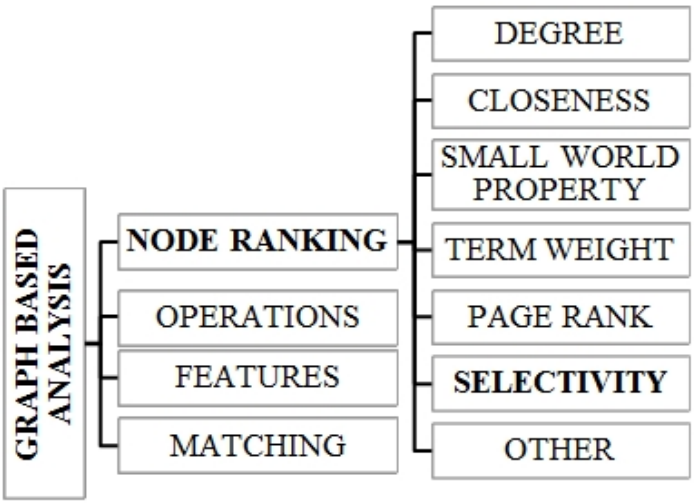


Figure 7.4: Classification of graph-based methods, modified from [116].

NP-chunks gives better precision than n-grams , and by adding the POS tag(s) to the terms improves the results independent of the term selection approach applied.

Turney (2003) in [119] implements enhancements to the Kea keyphrase extraction algorithm by using statistical associations between keyphrases and enhances the coherence of the extracted keywords.

Song et al. (2003) represent the Information Gain-Based keyphrase extraction system called KPSpotter [117].

HaCohen-Kerner et al. (2005) in [88] investigate the automatic extraction and learning of keyphrases from scientific articles written in English. They use various machine learning (ML) methods and report that the best results are achieved with J48 (an improved variant of C4.5).

Medelyan and Witten (2006) propose a new method called KEA++, which enhances automatic keyphrase extraction by using semantic information on terms and phrases gleaned from a domain-specific thesaurus [102]. KEA++ is actually an improved version of the previously mentioned Kea devised by Witten et al. Zhang Y. et al.

The group of researchers in [130] (2006) propose the use of not only 'global context information', but also 'local context information'. For the task of keyword extraction they engage Support Vector Machines (SVM). Experimental results indicate that the proposed SVM based method can significantly outperform the baseline methods for keyword extraction. Wang (2006) in [123] exploits different text features in order to determine whether a phrase is a keyphrase: TF and IDF, appearance in the title or headings (subheadings) of the given document, and the frequency appearing in the paragraphs of the given document in the combination with Neural Networks are proposed.

Nguyen and Kan (2007) [108] propose an algorithm for keyword extraction from scientific publications using linguistic knowledge. They introduce features that capture salient morphological phenomena found in scientific keyphrases, such as whether a candidate keyphrase is an acronym or weather it uses specific terminologically productive suffixes.

Zhang C. et al. (2008) in [129] implement a keyword extraction method from documents using Conditional Random Fields (CRF). The CRF model is a state-of-the-art sequence labeling method, which can use the features of documents more sufficiently and efficiently, and considers the keyword extraction as the string labeling task. The CRF model outperforms other ML methods such as SVM, Multiple Linear Regression model, etc.

Krapivin et al. (2010) in [95] use NLP techniques to improve various ML approaches (SVM, Local SVM, Random Forests) to the task of automatic keyphrase extraction from scientific papers. Evaluation shows promising results that outperform state-of-the-art Bayesian learning system KEA on the same dataset without the use of controlled vocabularies.

7.5.2 Unsupervised

HaCohen-Kerner (2003) in [89] presents a simple model that extracts keywords from abstracts and titles. The model uses unigrams, 2-grams and 3-grams, and a stopwords[2] list. The highest weighted group of words (merged and sorted n-grams) is proposed as keywords.

Pasquier (2010) in [112] describes the design of a keyphrase extraction algorithm for a single document using sentence clustering and Latent Dirichlet Allocation. The principle of the algorithm is to cluster sentences of the documents in order to highlight parts of text that are semantically related. The clustering is performed by using the cosine similarity between sentence vectors, K-means, Markov Cluster Process (MCP) and ClassDens techniques. The clusters of sentences, that reflect the themes of the document, are analyzed for obtaining the main topic of the text. The most important words from these topics are proposed as keyphrases.

Pudota et al. (2010) in [113] design a domain independent keyphrase extraction system that can extract potential phrases from a single document in an unsupervised, domain-independent way. They engaged n-grams, but they also incorporated linguistic knowledge (POS tags) and statistics (frequency, position, lifespan) of each n-gram in defining candidate phrases and their respective feature sets.

Hurt in [93] examines the differences between author generated keywords and automatically generated keywords using an inverse frequency and maximum likelihood algorithm. They express results in terms of novel linguistic measure 'keyness', which is defined as a log-likelihood measure of the relatedness of one or more specified words, keywords, to a corpus of literature. Testing of these two methods, they show that there are no statistically significant differences in the achieved results.

Very recent research by Yang et al. (2013) [128] focus on keyword extraction based on entropy difference between the intrinsic and extrinsic modes, which refers to the fact that relevant words significantly reflect the author's writing intention. Their method uses the Shannon's entropy difference between the intrinsic and extrinsic mode, which refers to the occurrences of words as being modulated by the author's purpose, while the irrelevant words are distributed randomly in the text. They indicate that the ideas of this work can be applied to any natural language without requiring any previous knowledge semantics or syntax of the language, especially for single documents of which there is no a priori information available.

7.5.3 Graph-Based

Ohsawa et al. (1998) in [109] propose an algorithm for the automatic indexing by co-occurrence graphs constructed from metaphors, called KeyGraph. This algorithm is based on the segmenting of a graph, representing the co-occurrence between terms in a document, into clusters. Each cluster corresponds to a concept on which the author's idea is based, and top ranked terms by a statistic based on each term's relationship to these clusters are selected as keywords. KeyGraph proved to be a content sensitive, domain independent device of indexing.

Matsou et al. (2001) in [101] present early research where a text document is represented as an undirected and unweighted co-occurrence network. Based on the network topology, the authors proposed an indexing system called KeyWorld, which extracts important terms (pairs of words) by measuring their contribution to small-world properties. The contribution of the vertex is based on

[2] Stopwords are the most frequent function words, which do not carry strong semantic properties, but are needed for the syntax of the language.

the closeness centrality calculated as the difference in small-world properties of the network with the temporarily elimination of a vertex combined with the inverse document frequency (idf).

Erkan and Radev (2004) in [85] introduce a stochastic graph-based method for computing the relative importance of textual units on the problem of text summarization by extracting the most important sentences. LexRank calculates sentence importance based on the concept of the eigenvector centrality in a graphical representation of sentences. A connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graphical representation of sentences. LexRank is shown to be quite insensitive to the noise in the data.

Mihalcea and Tarau (2004) in [106] report upon a seminal research which introduced a state-of-the-art TextRank model. TextRank is derived from PageRank and introduced to graph based text processing, keyword and sentence extraction tasks. The abstracts are modeled as undirected or directed and weighted co-occurrence networks using a co-occurrence window of variable sizes (2-10). The lexical units are preprocessed: stopwords removed, words restricted with POS syntactic filters (open class words, nouns and adjectives, nouns). The PageRank motivated score of the importance of the vertex derived from the importance of the neighboring vertices is used for keyword extraction. The obtained TextRank performance compares favorably with the supervised machine learning n-gram based approach.

Mihalcea (2004) in [104] presents an extension to earlier work [106], where the TextRank algorithm is applied for the text summarization task powered by sentence extraction. In this task TextRank performed on a par with the supervised and unsupervised summarization methods, which motivated the new branch of research based on the graph-based extracting and ranking algorithms .

Xie (2005) in [127] studies different centrality measures in order to predict noun phrases that appear in the abstracts of scientific articles. The measures tested are: degree, closeness, betweenness and information centrality . Their results show that centrality measures improve the accuracy of the prediction in terms of both precision and recall. Furthermore, the method of constructing a noun-phrase (NP) network significantly influences the accuracy when using the centrality heuristic itself, but is negligible when it is used together with other text features in decision trees.

Huang et al. (2006) [91] propose an automatic keyphrase extraction algorithm using an unsupervised method also based on connectedness and betweenness centrality.

Palshikar (2007) in [110] proposes a hybrid structural and statistical approach to extract keywords from a single document . The undirected co-occurrence network, using a dissimilarity measure between two words, calculated from the frequency of their co-occurrence in the preprocessed and lemmatized document, as the edge weight, was shown to be appropriate for the centrality measures based approach for keyword extraction.

Wan and Xiao (2008) in [122] propose a small number of nearest neighbor documents to provide more knowledge to improve single document keyphrase extraction. A specified document is expanded to a small document set by adding a few neighbor documents close to the document using a cosine similarity measure, while the term weight is computed by TF-IDF . The local information in the specified document and the global information in all the neighboring documents are taken into consideration along with the expanded document set using a graph-based ranking algorithm .

Litvak and Last (2008) in [98] compare supervised and unsupervised approaches for keywords identification in the task of extractive summarization. The approaches are based on the graph-based syntactic representation of text and web documents. The results of the HITS algorithm on a set of summarized documents performed comparably to supervised methods (Naïve Bayes, J48, SVM). The authors suggest that simple degree-based rankings from the first iteration of HITS, rather than running it to its convergence, should be considered.

Grineva et al. (2009) in [87] use community detection techniques for the extraction of key terms on Wikipedia's texts, modeled as a graph of semantic relationships between terms. The

results show that the terms related to the main topics of the document tend to form a community, thematically cohesive groups of terms. Community detection allows the effective processing of multiple topics in a document and efficiently filters out noise. The results achieved on weighted and directed networks from semantically linked, morphologically expanded and disambiguated n-grams from the articles' titles. Additionally, for the purpose of testing noise stability, they repeated the experiment on different multi-topic web pages (news, blogs, forums, social networks, product reviews) which confirmed that community detection outperforms TF-IDF model.

Tsatsaronis et al. (2010) in [118] present SemanticRank, a network based ranking algorithm for keyword and sentence extraction from text. Semantic relation is based on the calculated knowledge-based measure of semantic relatedness between linguistic units (keywords or sentences). The keyword extraction from the Inspec abstracts' results reported a favorable performance of SemanticRank over state-of-the-art counterparts - weighted and unweighted variations of PageRank and HITS.

Litvak et al. (2011) in [99] introduce DegExt, a graph-based language independent keyphrase extractor, which extends the keyword extraction method described in [98]. They also compare DegEx with state-of-the-art approaches: GenEx [121] and TextRank [106]. DegEx surpasses both in terms of precision, implementation simplicity and computational complexity.

Boudin (2013) in [80] compares various centrality measures for graph-based keyphrase extraction. Experiments on standard data sets of English and French show that simple degree centrality achieves results comparable to the widely used TextRank algorithm; and that closeness centrality obtains the best results on short documents. Undirected and weighted co-occurrence networks are constructed from syntactically (only nouns and adjectives) parsed and lemmatized text using a co-occurrence window. Degree, closeness, betweenness and eigenvector centrality are compared to the PageRank motivated method proposed by Mihalcea (2004) in [106] as a baseline. Degree centrality achieves a similar performance as the much more complex TextRank. Closeness centrality outperforms TextRank on short documents (scientific papers abstracts).

Zhou et al. (2013) in [132] investigate a weighted complex network based keyword extraction incorporating the exploration of the network structure and linguistics knowledge. The focus is on the construction of a lexical network including the reasonable selection of vertices, the proper description of the relationships between words, a simple weighted network and TF-IDF. The reasonable selection of words from texts as lexical vertices from a linguistic perspective, the proper description of the relationship between words and the enhancement of vertex attributes attempt to represent texts as lexical networks more accurately. The Jaccard coefficient is used to reflect the associations or relationships of two words rather than the usual co-occurrence criteria in the process of network construction. The importance of each vertex to become a keyword candidate is calculated with closeness centrality. The compound measures that takes vertex's attributes (words length and IDF) are applied. Approach is compared with three competitive baseline approaches: binary network, simple weighted network and TF-IDF approach. Experiments for Chinese indicate that the lexical network constructed by this approach achieves preferable effect on accuracy, recall and F-score over the classic TF-IDF method.

Lahiri et al. (2014) in [96] extract keywords and keyphrases from co-occurrence networks of words and from noun-phrases collocations' networks. Eleven measures (degree, strength, neighborhood size, coreness, clustering coefficient, structural diversity index, page rank, HITS hub and authority score , betweenness, closeness and eigenvector centrality) are used for keyword extraction from directed/undirected and weighted networks. The obtained results from four data sets suggest that centrality measures outperform the baseline term frequency - inverse document frequency (TF-IDF) model, and simpler measures such as degree and strength outperform computationally the more expensive centrality measures such as coreness and betweenness.

Abilhoa and de Castro (2014) in [75] propose a keyword extraction method representing tweets

(microblogs) as graphs and apply centrality measures for finding the relevant keywords. They developed a technique named Twitter Keyword Graph where in the pre-processing step they use tokenization, stemming and stopwords' removal. Keywords are extracted from the graph cascade-like applying graph centrality measures - closeness and eccentricity. The performance of the algorithm is tested on a single text from the literature and compared with the TF-IDF approach and KEA algorithm. Finally, the algorithm is tested on five sets of tweets of increasing size. The computational time to run the algorithms proved to be a robust proposal to extract keywords from texts, especially from short texts such as microblogs.

Beliga et al. (2014) in [59] propose the selectivity-based keyword extraction (SBKE) as a new unsupervised method for network-based keyword extraction. This approach is built with a new network measure - the vertex selectivity (defined as the average weight distribution on the edges of the single vertex) - see Section 7.5. In [59] is also shown that selectivity slightly outperforms the standard centrality-based measures: in-degree, out-degree, betweenness and closeness. Vertices with the highest selectivity value are open-class words (content words) which are preferred keyword candidates (nouns, adjectives, verbs) or even part of collocations, keyphrases, names, etc. Selectivity is insensitive to non-content words or stopwords and therefore can efficiently detect semantically rich open-class words from the network and extract keyword candidates.

Centrality measures are discriminative properties of the importance of a vertex in a graph, and are directly related to the structure of the graph [75]. The Table 7.1 in parts one and two overviews network measures that are widely used in graph-based research on keyword extraction, together with additional measures from the NLP domain. Mark asterisk (*) denotes graph-based measures.

NAME	DEFINITION	RESEARCH
Degree*	Number of edges incident to a vertex.	[59, 80, 96, 127]
Strength*	Sum of the weights of the edges incident to a vertex.	[96]
Selectivity*	Fraction of the vertex strength and vertex degree (average strength).	[59]
Neighborhood size*	Number of immediate neighbors to a vertex.	[96]
Coreness*	Outermost core number of a vertex in the k-core decomposition of a graph.	[96]
Clustering Coefficient*	Density of edges among the immediate neighbors of a vertex.	[96]
Page Rank*	Importance of a vertex based on how many important vertices it is connected to.	[96, 118]
TextRank*	Modification of an algorithm derived from Google's PageRank is based upon the eigenvector centrality measure and implement the concept of 'voting'.	[80, 104]
HITS*	Importance of a vertex as a hub (pointing to many others) and as an authority (pointed to by many others).	[96, 98, 118]
Betweenness*	The fraction of shortest paths that pass through a vertex, calculated over all vertex pairs - the measure of how many shortest paths between all other node-pairs are traversing a node.	[59, 80, 91, 96, 127]
Closeness*	Reciprocal of the sum of distances of all vertices to some vertex.	[59, 75, 80, 96, 101, 127, 132]
Community detection*	Community detection techniques are based on the principles which detect nodes with dense internal connections and sparser connections between groups.	[87]
Eigenvector Centrality*	Element of the first eigenvector of a graph adjacency matrix corresponding to a vertex.	[80, 96]
Information Centrality	Generalization of betweenness centrality - focuses on the information contained in all paths originating with a specific actor.	[127]
Structural Diversity Index	Normalized entropy of the weights of the edges incident to a vertex.	[96]
The Jaccard coefficient or Jaccard index	Reflects the association or relationship of two words taking into account not only the co-occurrence frequency, but also the frequency of both words in a pair.	[132]

Table 7.1: Measures and algorithms used for keyword extraction - Part 1 (asterisk (*) denotes graph-based measures).

NAME	DEFINITION	RESEARCH
Information Gain	The Kullback-Leibler divergence - a measure of expected reduction in entropy based on the 'usefulness' of an attribute.	[117]
TF, IDF, TF-IDF	Term frequency, inverse document frequency.	[87,93,96,101,110] [113,121–123,125, 132]
n-gram	N-gram is a contiguous sequence of n items from a given sequence of text or speech.	[89,92,106,113]
Cosine similarity	Determines similarity between two vectors.	[85,112,122]
SingleRank	Compute word scores for each single document based on the local graph for the specified document.	[122]
ExpandRank	Compute word scores for each single document based on the neighborhood knowledge of other documents.	[122]
Shannon's entropy difference	The difference between the intrinsic and extrinsic entropy.	[128]
Keyphraseness	The linear combination of features: phrase frequency, pos value, phrase depth, phrase last occurrence, phrase lifespan.	[100]
Other	Harmonic centrality, LIN centrality, Katz centrality, Wiener index, eccentricity, connectedness [127], POS tags [92,113], CRF [129], LexRank [85], SemanticRank [118], SimRank, etc.	

Table 7.2: Measures and algorithms used for keyword extraction - Part 2.

7.6 Selectivity-Based Keyword Extraction

7.6.1 Dataset

For the network based keyword extraction we use the data set composed of Croatian news articles [107]. The data set contains 1020 news articles from the Croatian News Agency (HINA), with manually annotated keywords (key phrases) by human experts. The set is divided as such: 960 annotated documents for learning of supervised methods, and 60 documents for testing. The test set of 60 documents is annotated by 8 different experts. We selected the first 30 texts from HINA's collection for our experiment.

7.6.2 Co-occurrence Network Construction

Each text can be represented as a complex network of linked words: each individual word is a vertex and the interactions amongst words are edges. Co-occurrence networks exploit simple neighbor relation; two words are linked if they are adjacent in the sentence [99]. The weight of the edge is proportional to the overall co-occurrence frequencies of the corresponding word pairs within a corpus. From the documents in the HINA data set we construct directed and weighted co-occurrence networks: one from the text in each document .

7.6.3 Results

We compute centrality measures for each vertex in a network constructed from 60 news articles: in-degree, out-degree, closeness, betweenness and selectivity. Then we rank all vertices (words) according to the values of each of these measures, obtaining the top 15 keyword candidates automatically from the network. It is obvious that top 15 ranked words according to the in/out degree centrality, closeness centrality and betweenness centrality are stopwords (conjunctions, prepositions, determiners, etc.) - see Table 7.3. It can also be noticed that centrality measures return almost identical stopwords. However, the selectivity measure ranked only open-class words: nouns, verbs and adjectives. We expect that among these highly-ranked words are keyword candidates. The same results are shown in the preliminary research on keyword extraction from multitopic web documents [72].

IN-DEGREE	OUT-DEGREE	CLOSENESS	BETWEENNESS	SELECTIVITY
biti (is/be)	biti (is/be)	biti (is/be)	biti (is/be)	Bratislava
i(and)	i (and)	i (and)	i (and)	području (area)
u (in)	u (in)	taj (that/this)	u (in)	utorak (Tuesday)
a (but/and)	a (but/and)	na (on)	a (but/and)	zaled'e (hinterland)
da (that/to)	sebe (self)	sebe (self)	sebe (self)	revolucije (revolution)
koji (which)	za (for)	on (he)	da (that/to)	provjera (check)
a (for)	taj (that/this)	da (that/to)	taj (that/this)	II. (roman number)
a (but/and)	da (that/to)	u(in)	koji (which)	desetljeća (decades)
taj (that/this)	od (from)	ali (but)	za (for)	Balkanu (Balkan)
sebe (self)	s (with)	za (for)	hrvatski (Croatian)	sloboda (freedom)
s (with)	a (but/and)	kako (how)	a (but/and)	universe
od (of)	koji (which)	hrvatski (Croatian)	od (from)	trophy
ne (not/no)	ne (not/no)	još (more/yet)	s (with)	stotina(hundred)
hrvatski (Croatian)	hrvatski (Croatian)	sad (now)	ne (not/no)	Splitu (Split)
o (on/about)	će (will)	godina (year)	iz (from)	razlika (difference)

Table 7.3: The top 15 ranked words according to the measures: in-degree, out-degree, closeness, betweenness and selectivity from th whole HINA dataset.

In short, it seems that selectivity is insensitive to stopwords and therefore can efficiently detect semantically rich open-class words from the network and extract better keyword candidates (which are probably names, parts of collocations or key phrases).

Simple measures such as selectivity promulgates the views and opportunities for the development of new graph-based methods which can yield successful keyword ranking, and at the same time circumvent the usage of demanding NLP procedures, which are deeply rooted in standard KE techniques. If we take into consideration the complexity and computational resources, then it is clear that the graph-based methods may have the advantage over traditional supervised and unsupervised methods. This is the reason why it makes sense to continue the work towards developing new graph-based methods.

7.7 Conclusion and Future Trends

Keywords provide a compact representation of a document's content. Graph-based methods for keyword extraction are inherently unsupervised, and have the fundamental aim to build a network of words (phrases or linguistic units) and then rank the vertices exploiting the measures of the network structure, usually centrality motivated. This Chapter is a detailed systemization of existing

approaches for keyword extraction: the review of related work on supervised and unsupervised methods with a special focus on the graph-based methods. The Chapter consolidates the most commonly used centrality measures that are essential in graph-based methods: in/out-degree, closeness, betweenness, etc. In addition, the existing work of Croatian extraction is included as well.

This work provides an insight into the related work of graph-based keyword extraction methods which successfully consolidated various techniques of natural language processing and complex network analysis. Combinations of these techniques establish a solid platform regard to the objectives of keyword (term) extraction and scope of the specific application. Such hybrid techniques represent new convenient ways to circumvent anomalies that occur in VSM and other traditionally used models.

Graph-based methods for keyword extraction are simple and robust in many ways: (1) they do not require advanced linguistic knowledge or processing, (2) they are domain independent and (3) they are language independent. Such graph-based KE techniques are certainly applicable for various tasks: text classification, summarization, search, etc. Due to the aforementioned benefits it is reasonable to expect that graph-based extraction will attract the attention of the research community in the future. It can be expected that many text and document analyses will incorporate graph-based keyword extraction.

8. Network-based Keyword Extraction from Multitopic Web Documents

8.1 Abstract

In this work we analyse the selectivity measure calculated from the complex network in the task of the automatic keyword extraction. Texts, collected from different web sources (portals, forums), are represented as directed and weighted co-occurrence complex networks of words. Words are nodes and links are established between two nodes if they are directly co-occurring within a sentence. We test different centrality measures for ranking nodes - keyword candidates. The promising results are achieved using the selectivity measure. Then we propose an approach which enables extracting word pairs according to the values of the in/out-selectivity and weight measures combined with filtering.

8.2 Introduction

Keyword extraction is an important task in the domain of the Semantic Web development. It is a problem of automatic identification of the important terms or phrases in text documents. It has numerous applications: information retrieval, automatic indexing, text summarization, semantic description and classification, etc. In the case of web documents it is a very demanding task: it requires extraction of keywords from web pages that are typically noisy, overburden with information irrelevant to the main topic (navigational information, comments, future announcements, etc.) and they usually contain several topics [87]. Therefore, in keyword extraction from web pages we are dealing with noisy and multitopic datasets .

Various approaches have been proposed for keywords and keyphrases identification (extraction) task. There are two main classes of approaches: supervised and unsupervised. Supervised approaches are based on using machine learning techniques on the manually annotated data [120, 125]. Therefore supervised approaches are time consuming and expensive. Unsupervised approaches may include clustering [133], language modelling [137] and graph-based approaches. Unsupervised approaches may also require different sets of external data, however these approaches are not depended on manual annotation. These approaches are more robust, but usually less

precise [80] .

A class of graph-based keyword extraction algorithms overcome some of these problems. In graph-based or network-based approaches the text is represented as a network in a way that words are represented as nodes and links are established between two nodes if they are co-occurring within the sentence. The main idea is to use different centrality measures for ranking nodes in the network. Nodes with the highest rank represent words that are candidates for keywords and keyphrases. In [96] an exhaustive overview of network centrality measures usage in the keyword identification task is given.

One of the probably most influential graph-based approaches is the TextRank ranking model introduced by Mihalcea and Tarau in [106]. TextRank is a modification of PageRank algorithm and the basic idea of this ranking technique is to determine the importance of a node according to the importance of its neighbours, using global information recursively drawn from the entire network. However, some recent researches have shown that even simpler centrality measures can give satisfactory results. Boudin [80] and Lahiri et al. [96] compare different centrality measures for keyword extraction task. Litvak and Last [98] compare supervised and unsupervised approach for keywords identification in the task of extractive summarization .

We have already experimented with graph-based approaches for Croatian texts representation. In [134, 135] we described graph-based word extraction and representation from the Croatian dictionary. We used lattice to represent different semantic relations (partial semantic overlapping, more specific, etc.) between words from the dictionary.

In [21, 50, 71] we described and analysed network-based representation of Croatian texts.

In [50] our results showed that in-selectivity and out-selectivity values from shuffled texts are constantly below selectivity values calculated from normal texts. It seems that selectivity measure is able to capture typical word phrases and collocations which are lost during the shuffling procedure. The same holds for English where Masucci and Rodgers [19] found that selectivity somehow captures the specialized local structures in nodes' neighborhood and forms of the morphological structures in text. According to these results, we expected that node selectivity may be potentially important for the text categories differentiation and include it in the set of standard network measures. In [71] we show that the node selectivity measure can capture structural differences between two genres of text.

This was the motivation for further exploration of selectivity for keyword extraction task from Croatian multitopic web documents. We have already analysed the selectivity-based keyword extraction in Croatian news [59]. In this Chapter we propose an in/out-selectivity based approach combined with filtering to extract keyword candidates from the co-occurrence complex network of text. We design selectivity-based approach as unsupervised, data and domain independent. In its basic form, only the stopwords list is a prerequisite for applying stopwords-filter. As designed, it is a very simple and robust approach appropriate for extraction from large multitopic and noisy datasets.

In Section 8.3 we present measures for the network structure analysis. In Section 8.4 we describe datasets and the construction of co-occurrence networks from used text collection. In Section 8.5 are the results of keyword extraction, and in the final Section 8.6, we elaborate the obtained results and make conclusions regarding future work.

8.3 The Network Measures

This Section describes basic network measures that are necessary for understanding our approach. More details about these measures can be found in [19, 67, 136]. In the network, N is the number of nodes and K is the number of links. In weighted language networks every link connecting two nodes i and j has an associated weight w_{ij} that is a positive integer number.

The node degree k_i is defined as the number of links incident upon a node. The in degree and out degree $k_i^{in/out}$ of node i is defined as the number of its in and out neighbours.

Degree centrality of the node i is the degree of that node. It can be normalised by dividing it by the maximum possible degree $N-1$:

$$\mathrm{dc}_i = \frac{k_i}{N-1}. \tag{8.1}$$

Analogously, the in-degree centralities are defined as in-degree of a node :

$$\mathrm{dc}_i^{in} = \frac{k_i^{in}}{N-1}. \tag{8.2}$$

The out-degree centrality of a node is defined in a similar way . Closeness centrality is defined as the inverse of farness, i.e. the sum of the shortest paths between a node and all the other nodes. Let d_{ij} be the shortest path between nodes i and j. The normalised closeness centrality of a node i is given by :

$$\mathrm{cc}_i = \frac{N-1}{\sum_{i\neq j} d_{ij}}. \tag{8.3}$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Let σ_{jk} be the number of shortest paths from node j to node k and let $\sigma_{jk}(i)$ be the number of those paths that traverse through the node i. The normalised betweenness centrality of a node i is given by :

$$\mathrm{bc}_i = \frac{\sum_{i\neq j\neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}}{(N-1)(N-2)}. \tag{8.4}$$

The strength of a node i is a sum of weights of all links incident with the node i :

$$\mathrm{s}_i = \sum_j w_{ij}. \tag{8.5}$$

All given measures are defined for directed networks, but language networks are weighted, therefore, the weights should be considered. In the directed network, the in-strength s_i^{in} of the node i is defined as the number of its incoming links, that is :

$$\mathrm{s}_i^{in} = \sum_j w_{ji}. \tag{8.6}$$

The out-strength is defined in a similar way . The selectivity measure is introduced in [19]. It is actually an average strength of a node. For a node i the selectivity is calculated as a fraction of the node weight and node degree :

$$\mathrm{e}_i = \frac{\mathrm{s}_i}{k_i}. \tag{8.7}$$

In the directed network, the in-selectivity of the node i is defined as :

$$\mathrm{e}_i^{in} = \frac{\mathrm{s}_i^{in}}{k_i^{in}}. \tag{8.8}$$

The out-selectivity is defined in a similar way .

Dataset	GL	NN	IN	SD
Number of words	199 417	146 731	118 548	44 367
Number of nodes N	27727	13036	15065	9553
Number of links K	105171	55661	28972	25155

Table 8.1: The number of words, number of nodes and number of links for all 4 datasets.

8.4 Methodology

8.4.1 The Construction of Co-occurrence Networks

Dataset contains 4 collections of web documents written in Croatian language collected from different web sources (portals and forums on different daily topics). The 4 different web sources: business portal Gospodarski list (GL), legislative portal Narodne novine (NN), news portal with forum Index.hr (IN), daily newspaper portal Slobodna Dalmacija (SD). The first step in networks construction was text preprocessing: "cleaning" special symbols, normalising Croatian diacritics (č, ć, ž, š, dž), and removing punctuations which does not mark the end of a sentence. Commonly, for Croatian which is highly flective Slavic language the lemmatisation and part-of-speech tagging should be performed, but we model our experiment without any explicit language knowledge.

For each dataset we constructed weighted and directed co-occurrence network. Nodes are words that are linked if they are direct neighbours in a sentence. The next step was introducing the networks as weighted edgelists, which contain all the pairs of connected words and their weights (the number of connections between two same words). In the Table 8.1 there are number of words, number of nodes and number of links per each dataset. We used Python and the NetworkX software package developed for the construction, manipulation, and study of the structure, dynamics, and functions of complex networks [8].

8.4.2 The Selectivity-based Approach

The goal of this experiment is to analyse the selectivity measure in the automatic keyword extraction task. First, we compute centrality measures for each node in all 4 networks: in-degree centrality, out-degree centrality, closeness centrality, betweenness centrality and selectivity centrality. Then we rank all nodes (words) according to the values of each of these measures, obtaining top 10 keyword candidates automatically from the network.

In the second part of our experiment we compute in-selectivity and out-selectivity for each node in all 4 networks. The nodes are then ranked according to the highest in/out-selectivity values. Then, for every node we detect neighbour nodes with the highest weight. For the in-selectivity we isolate one neighbour node with the highest outgoing link weight. For the out-selectivity we isolate one neighbour node with the highest ingoing link weight. The result of in/out-selectivity extraction is a set of ranked word tuples.

The third part of our approach consider applying different filters on the in/out-selectivity based word tuples. The first is the stopwords-filter: we filter out all tuples that contain stopwords . 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 second is the high-weights-filter: from the in/out-selectivity based word tuples we chose only those tuples that have the same values for the selectivity and weight. The third filter is the combination of the first two filters.

	selectivity	in-degree	out-degree	closeness	betweenness
1.	mladićevi (joungsters)	i (and)	i (and)	je (is)	i (in)
2.	pomlatili (beaten)	u (in)	je (is)	i (and)	je (is)
3.	seksualnog (sexual)	je (is)	u (in)	se (self)	u (in)
4.	policijom (police)	na (on)	na (on)	da (that)	na (on)
5.	uhićeno (arrested)	da (that)	se (self)	su (are)	se (self)
6.	skandala (scandal)	za (for)	za (for)	to (it)	za (for)
7.	podnio (submitted)	se (self)	su (are)	a (but)	da (that)
8.	obožavatelji (fans)	a (but)	da (that)	će (will)	su (are)
9.	sata (hour)	su (are)	s (with)	samo (only)	a (but)
10.	Baskiji (Baskia)	s (with)	od (from)	ali (but)	s (with)

Table 8.2: Top ten words from the dataset IN ranked according to the selectivity, in/out-degree, closeness and betwenness.

8.5 Results

Initially, we analyse 4 networks constructed for each dataset. The top 10 ranked nodes with the highest values of the selectivity, in degree, out degree, closeness and betwenness measures for datasets IN, GL, SD and NN are shown in the Tables 8.2, 8.3, 8.4 and 8.5. It is obvious that top 10 ranked words according to the in/out degree centrality, closeness centrality and betwenness centrality are stopwords. It can be also noticed that centrality measures return almost identical top 10 stopwords. However, the selectivity measure ranked only open-class words: nouns, verbs and adjectives. We expect that among these highly ranked words are keyword candidates.

Furthermore, we analyse selectivity measure in details. Since texts are better represented as directed networks [18], we analyse words with in-selectivity and out-selectivity measure separately. We extract word-tuple: the word before for in-selectivity and the word after for out-selectivity that has the highest value of the weight. In Table 8.6 are shown ten highly ranked in/out-selectivity based word-tuples together with the values of in/out-selectivity and weight.

Hence, we extract the most frequent word-tuples which are possible collocations or phrases from the text. We expect that among these highly ranked word-tuples are keyword candidates. Due to limited space, we show results only for the NN dataset, but other datasets raised similar results.

In Table 8.6 there are word-tuples which contain stopwords, especially for the in-selectivity based ranking.Therefore we use stopwords-filter defined in the previous Section as shown in Table 8.7. Now we obtain more open class keyword candidates from highly ranked word-tuples.

In Table 8.8. there are 10 highly ranked word-tuples for the NN dataset with the high-weights-filter applied. Using this approach some new keyword candidates appear in the ranking results.

In Table 8.9. there are 10 highly ranked word-tuples from the NN dataset with the both filters applied. According to our knowledge about the content of the dataset, these two filters derived the best results.

8.6 Conclusion and Discussion

We analyse network-based keyword extraction from multitopic Croatian web documents using selectivity measure. We compare keyword candidate words rankings with selectivity and three network-based centrality measures (degree, closeness and betwenness). The selectivity measure gives better results because centrality-based rankings select only stopwords as the top 10 ranked words. Furthermore, we propose extracting the highly connected word-tuples with the highest in/out-selectivity values as the keyword candidates. Finally, we apply different filters (stopwords-filter,

	selectivity	in degree	out degree	closeness	betweenness
1.	stupastih (cage)	i (and)	i (and)	i (and)	i (and)
2.	populaciju (population)	u (in)	u (in)	se (self)	u (in)
3.	izdanje (issue)	na (on)	je (is)	je (is)	je (is)
4.	online (online)	je (is)	se (self)	su (are)	na (on)
5.	webshop (webshop)	ili (or)	na (on)	a (but)	se (self)
6.	matrica (matrix)	a (but)	ili (or)	ili (or)	ili (or)
7.	pretplata (subscription)	se (self)	su (are)	to (it)	a (but)
8.	časopis (journal)	za (for)	za (for)	bolesti (disease)	za (for)
9.	oglasi (ads)	od (from)	od (from)	da (that)	su (are)
10.	marketing (marketing)	su (are)	a (but)	biljke (plants)	od (from)

Table 8.3: Top ten words from the dataset GL ranked according to the selectivity, in/out-degree, closeness and betwenness.

	selectivity	in-degree	out-degree	closeness	betweenness
1.	seronjo (bullshitter)	i (and)	i (and)	i (and)	i (and)
2.	Splitu (Split)	u (in)	je (is)	je (is)	je (is)
3.	upišite (fill-in)	je (is)	u (in)	svibanj (May)	u (in)
4.	uredniku (editor)	komentar (comment)	se (self)	se (self)	se (self)
5.	ekrana (monitor)	na (on)	svibanj	ali (but)	na (on)
6.	crkvu (church)	se (self)	na (on)	a (but)	od (from)
7.	supetarski (Supetar)	za (for)	za (for)	će (will)	za (for)
8.	vijesti (news)	a (but)	da (that)	to (it)	a (but)
9.	zaradom (earning)	svibanj (May)	ne (ne)	još (more)	svibanj
10.	Jović (Jović)	od (from)	a (but)	pa (so)	to (it)

Table 8.4: Top ten words from the dataset SD ranked according to the selectivity, in/out-degree, closeness and betwenness.

	selectivity	in-degree	out-degree	closeness	betweenness
1.	novine (newspaper)	i (and)	i (and)	i (and)	i (and)
2.	temelju (based on)	u (in)	u (in)	ili (or)	u (in)
3.	manjinu (minority)	za (for)	je (is)	je (is)	za (for)
4.	srpsku (Serbian)	na (on)	za (for)	se (self)	ili (or)
5.	sladu (harmony)	ili (or)	se (self)	da (that)	na (on)
6.	snagu (strength)	iz (from)	ili (or)	usluga (service)	je (is)
7.	osiguranju (insurance)	te (and)	na (on)	zakona (law)	se (self)
8.	narodnim (national)	je (is)	o (on)	a (but)	o (on)
9.	novinama (newspaper)	se (self)	te (and)	skrbi (welfare)	te (and)
10.	kriza (crisis)	s (with)	članak (article)	HRT-a (HRT-a)	iz (form)

Table 8.5: Top ten words from the dataset NN ranked according to the selectivity, in/out-degree, closeness and betwenness.

	in-selectivity			out-selectivity		
	word tuple	e^{in}	w	word tuple	e^{out}	w
1.	narodne **novine**	326	326	**srpsku** nacionalnu	222	222
2.	na **temelju**	317	317	**nacionalnu** pripadnost	183	1
3.	nacionalnu **manjinu**	275	2	**ovjesne** jedrilice	159	159
4.	za **srpsku**	222	222	**narodnim** novinama	129	129
5.	u **skladu**	202	202	**narodne** jazz	111	1
6.	na **snagu**	172	172	**manjinu** gradu	78	1
7.	o **osiguranju**	134	43	**ovoga** sporazuma	72	1
8.	u **narodnim**	129	129	**crvenog** kristala	72	3
9.	narodnim **novinama**	129	129	**skladu** provjeriti	67	1
10.	crvenog **križa**	99	2	**oružani**h sukoba	58	4

Table 8.6: Top ten highly ranked in/out-selectivity based word-tuples for the NN dataset.

	in-selectivity			out-selectivity		
	word tuple	e^{in}	w	word tuple	e^{out}	w
1.	narodne **novine**	326	326	**srpsku** nacionalnu	222	222
2.	nacionalnu **manjinu**	275	2	**nacionalnu** pripadnost	183	1
3.	narodnim **novinama**	129	129	**ovjesne** jedrilice	183	1
4.	crvenoga **križa**	99	2	**narodnim** novinama	129	129
5.	jedinicama **regionalne**	65	1	**narodne** jazz	111	1
6.	nacionalne **manjine**	61	61	**manjinu** gradu	78	1
7.	rizika **snaga**	57	1	**ovoga** sporazuma	72	1
8.	medije **ubroj**	47	1	**crvenog** kristala	72	3
9.	crveni **križ**	42	42	**skladu** provjeriti	67	1
10.	uopravni **spor**	41	41	**oružanih** sukoba	58	4

Table 8.7: Top ten highly ranked in/out-selectivity based word-tuples without stopwords for the NN dataset.

in-selectivity		out-selectivity	
word tuple	e^{in}=w	word tuple	e^{out}=w
na **temelju** (based on)	317	**ovjesne** jedrilice (hangh glider)	159
za **srpsku** (for Serbian)	222	**narodnim** novinama (Nat. news.)	129
u **skladu** (according to)	202	**sjedištem** u (headquarter in)	55
na **snagu** (into effect)	172	**objavit** će (will be bublished)	53
u **narodnim** (in national)	129	**republici** Hrvatskoj (Croatia)	52
narodnim **novinama** (Nat. news.)	129	**albansku** nacionalnu (Alb. nat.)	52
i **dopunama** (and amendments)	68	**republika** Hrvatska (Croatia)	49
nacionalne **manjine** (nat. minority)	61	**oplemenjivačkog** prava (noble law)	45
sa **sjedištem** (with headquarter)	55	**madjarsku** nacionalnu (Hung. nat.)	40

Table 8.8: Top ten highly ranked in/out-selectivity based word-tuples with equal in/out-selectivity and weight for the NN dataset.

in-selectivity word tuple	out-selectivity word tuple
narodne **novine** (National newspaper)	**srpsku** nacionalnu (Serbian national)
narodnim **novinama** (Nat. newspapers)	**ovjesne** jedrilice (hangh glider)
nacionalne **manjine** (nat. minority)	**narodnim** novinama (Nat. newspapers)
crveni **križ** (red cross)	**republici** hrvatskoj (Republic of Croatia)
upravni **spor** (administrative dispute)	**albansku** nacionalnu (Albanian national)
ovjesnom **jedrilicom** (hangh glider)	**republika** hrvatska (Republic of Croatia)
elektroničke **medije** (electronic media)	**oplemenjivačkog** prava (noble law)
nacionalnih **manjina** (national minority)	**madjarsku** nacionalnu (Hungarian nat.)
domovinskog **rata** (Homeland War)	**romsku** nacionalnu (Romany national)
Ivan **Vrljić** (Ivan Vrljić)	**nadzorni** odbor (supervisory board)

Table 8.9: Top ten highly ranked in/out-selectivity based word-tuples with equal in/out-selectivity and weight without stopwords for the NN dataset.

high-weights-filter) in order to keyword candidate list.

The first part of analysis can raise some considerations regarding the selectivity measure. The selectivity measure is important for the language networks especially because it can differentiate between two types of nodes with high strength values (which means words with high frequencies). Nodes with high strength values and high degree values would have low selectivity values. These nodes are usually stopwords (conjunctions, prepositions,...). On the other side, nodes with high strength values and low degree values would have high selectivity values. These nodes are possible collocations, keyphrases and names that appear in the texts. It seems that selectivity is insensitive to stopwords (which are the most frequent words) and therefore can efficiently detect semantically rich open class words from the network.

Furthermore, since we modelled multitopic datasets the keyword extraction task is even more demanding. From the results of this preliminary research it seems that the selectivity has a potential to extract keyword candidates without preprocessing (lemmatisation, POS tagging) from multitopic sources.

There are several drawbacks in this reported work: we did not perform the classical evaluation procedure because we did not have annotated data and we conducted analysis only on Croatian texts.

For the future work we plan to evaluate our results on different datasets in different languages. Furthermore, it seems promising to define an approach that can extract a sequence of three or four neighbouring words based on filtered word-tuples. We also plan to experiment with lemmatised texts. Finally, in the future we will examine the effect of noise to the results obtained from multitopic sources.

9. Toward Selectivity Based Keyword Extraction for Croatian News

9.1 Abstract

Our approach proposes a novel network measure - the node selectivity for the task of keyword extraction. The node selectivity is defined as the average strength of the node. Firstly, we show that selectivity based keyword extraction slightly outperforms the extraction based on the standard centrality measures: in-degree, out-degree, betweenness, and closeness. Furthermore, from the data set of Croatian news we extract keyword candidates and expand extracted nodes to word-tuples ranked with the highest in/out selectivity values. The obtained sets are evaluated on manually annotated keywords: for the set of extracted keyword candidates the average $F1$ score is 24.63%, and the average $F2$ score is 21.19%; for the exacted word-tuples candidates the average $F1$ score is 25.9% and the average $F2$ score is 24.47%. Selectivity based extraction does not require linguistic knowledge while it is purely derived from statistical and structural information of the network.

9.2 Introduction

The task of keyword extraction is to automatically identify a set of terms that best describe the document [106]. Automatic keyword extraction establishes a foundation for various natural language processing applications: information retrieval, the automatic indexing and classification of documents, automatic summarization, high-level semantic description, etc.

State-of-the-art keyword extraction approaches are based on statistical methods which require learning from hand-annotated data sets. In the last decade the focus of research has shifted toward unsupervised methods, mainly towards network or graph enabled keyword extraction. In a network enabled keyword extraction the document representation may vary from very simple (words are nodes and their co-occurrence is represented with links), or can incorporate very sophisticated linguistic knowledge like syntactic [17] or semantic relations [118]. Typically, the source (document, text, data) for keyword extraction is modelled with one network. This way, both the statistical properties (frequencies) as well as the structure of the source text are represented by a unique formal representation, hence a complex network .

A network (or graph, since the number of words in isolated documents is limited) enabled

keyword extraction exploits different measures for the task of identifying and ranking the most representative features of the source - the keywords . The keyword extraction powered by network measures can be on the node, network or subnetwork level. Measures on the node level are: degree, strength, centrality [96]; on the network level : coreness, clustering coefficient, PageRank motivated ranking score or HITS motivated hub and authority score [80, 98, 106]; on the subnetwork level : communities [87]. Most of the of the research was motivated with various centrality measures : degree, betweenness, closeness and eigenvector centrality [80, 96, 98, 101, 106, 110].

Our research aims at proposing a novel selectivity based method for the unsupervised keyword extraction from the network of Croatian texts. Since Croatian is a highly flective Slavic language, the source text usually needs a substantial preprocessing (lemmatization - morphological normalization, stopwords removal, part-of-speech (POS) annotation, morphosyntactic descriptions (MSD) tagging, etc.), we design our approach with little or no linguistic knowledge. A new network measure - the node selectivity, originally proposed by Masucci and Rodgers [11, 19] (that can distinguish a real from a shuffled one), is applied to automatic keyword extraction. Selectivity is defined as the average weight distribution on the links of the single node. In our previous work, the node selectivity measure performed in favour of the differentiation between original and shuffled Croatian texts [21, 50], and for the differentiation of blog and literature text genres [71]. In this work we explore the potential of the selectivity for the keyword extraction in the Croatian news articles. To the best of our knowledge, the node selectivity measure has not been applied to the keyword extraction task before .

Section 9.3 presents an overview of related work on automatic keyword extraction. In Section 9.4 we present the definition of the measures for the network structure analysis. In Section 9.5 we present the construction of co-occurrence networks from collection of used text. The methods used for network based keyword extraction are explained in Section 9.6. The evaluation of obtained keywords and results are in Section 9.7. In the final Section, we elaborate upon the selectivity method and make conclusions regarding future work.

9.3 Related Work

Lahiri et al. [96] extract keywords and keyphrases form co-occurrence networks of words and from noun phrases collocations networks. Eleven measures (degree, strength, neighbourhood size, coreness, clustering coefficient, structural diversity index, page rank, HITS – hub and authority score, betweenness, closeness and eigenvector centrality) are used for keyword extraction from directed/undirected and weighted networks. The obtained results on 4 data sets suggest that centrality measures outperform the baseline term frequency/inverse document frequency (tf-idf) model, and simpler measures like degree and strength outperform computationally more expensive centrality measures like coreness and betweenness .

Boudin [80] compares various centrality measures for graph-based keyphrase extraction. Experiments on standard data sets of English and French show that simple degree centrality achieves results comparable to the widely used TextRank algorithm; and that closeness centrality obtains the best results on short documents . Undirected and weighted co-occurrence networks are constructed from syntactically (only nouns and adjectives) parsed and lemmatized text using co-occurrence window. Degree, closeness, betweenness and eigenvector centrality are compared to PageRank ad proposed by Mihalcea in [106] as a baseline. Degree centrality achieve similar performance as much complex TextRank. Closeness centrality outperforms TextRank on short documents (scientific papers abstracts).

Litvak and Last [98] compare supervised and unsupervised approaches for keywords identification in the task of extractive summarization . The approaches are based on the graph-based syntactic representation of text and web documents. The results of the HITS algorithm on a set of summarized documents performed comparably to supervised methods (Naive Bayes, J48, Support

Vector Machines). The authors suggest that simple degree-based rankings from the first iteration of HITS, rather than running it to its convergence, should be considered.

Grineva et al. [87] use community detection techniques for key terms extraction on Wikipedia's texts , modelled as a graph of semantic relationships between terms . The results showed that the terms related to the main topics of the document tend to form a community , thematically cohesive groups of terms. Community detection allows the effective processing of multiple topics in a document and efficiently filters out noise. The results achieved on weighted and directed networks from semantically linked, morphologically expanded and disambiguated n-grams from the article's titles. Additionally, for the purpose of the noise stability, they repeated the experiment on different multi-topic web pages (news, blogs, forums, social networks, product reviews) which confirmed that community detection outperforms td-idf model.

Palshikar [110] proposes a hybrid structural and statistical approach to extract keywords from a single document. The undirected co-occurrence network, using a dissimilarity measure between two words, calculated from the frequency of their co-occurrence in the preprocessed and lemmatized document, as the edge weight, was shown to be appropriate for the centrality measures based approach for keyword extraction.

Mihalcea and Tarau [106] report a seminal research which introduced a state-of-the-art TextRank model. TextRank is derived from PageRank and introduced to graph based text processing, keyword and sentence extraction. The abstracts are modelled as undirected or directed and weighted co-occurrence networks using a co-occurrence window of variable sizes (2..10). Lexical units are preprocessed: stopwords removed, words restricted with POS syntactic filters (open class words, nouns and adjectives, nouns). The PageRank motivated score of the importance of the node derived from the importance of the neighboring nodes is used for keyword extraction. The obtained TextRank performance compares favorably with the supervised machine learning n-gram based approach.

Matsou et al. in [101] present an early research where a text document is represented as an undirected and unweighted co-occurrence network. Based on the network topology, the authors proposed an indexing system called KeyWorld, which extracts important terms (pairs of words) by measuring their contribution to small-world properties. The contribution of the node is based on closeness centrality calculated as the difference in small-world properties of the network with the temporarily elimination of a node combined with inverse document frequency (idf).

Erkan and Radev [85] introduce a stochastic graph-based method for computing the relative importance of textual units on the problem of text summarization by extracting the most important sentences. LexRank calculates sentence importance based on the concept of eigenvector centrality in a graph representation of sentences . A connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences. LexRank is shown to be quite insensitive to the noise in the data.

Mihalcea in [104] presents an extension to earlier work [106], where the TextRank algorithm is applied for the text summarization task powered by sentence extraction. On this task TextRank performed on a par with the supervised and unsupervised summarization methods, which motivated the new branch of research based on the graph-based extracting and ranking algorithms.

Tsatsaronis et al. [118] present SemanticRank, a network based ranking algorithm for keyword and sentence extraction from text. Semantic relation is based on the calculated knowledge-based measure of semantic relatedness between linguistic units (keywords or sentences). The keyword extraction from the Inspec abstracts' results reported a favorable performance of SemanticRank over state-of-the-art counterparts - weighted and unweighted variations of PageRank and HITS.

Huang et al. [91] propose an automatic keyphrase extraction algorithm using an unsupervised method based on connectedness and betweeness centrality.

9.3.1 Related Work on the Croatian Language

The keyphrase extraction for¸ the Croatian language has been addressed in both supervised [76] and unsupervised [77, 107, 114] settings. Ahel et al. [76] use a Naive Bayes classifier combined with tf-idf (term frequency/inverse document frequen-cy), [107] utilizes the part-of-speech (POS) and morphosyntactic description (MSD) tags filtering followed by tf-idf ranking, and [114] exploits the distributional semantics to build topically related word clusters, from which they extract keywords and expand them to keyphrases. Bekavac et al. [77] propose a genetic programming approach for keyphrases the extraction for the Croatian language on the same data set. GPKEX can evolve simple and interpretable keyphrase scoring measures that perform comparably to other machine learning methods for Croatian. Reported research on extraction of Croatian keywords use a data set composed of Croatian news articles from the Croatian News Agency (HINA), with hand annotated keywords by human experts.

9.4 The Complex Network Analysis

This Section describes the basic network measures that are necessary for understanding our approach. More details about these measures can be found in [19, 67]. In the network, N is the number of nodes and K is the number of links. In weighted language networks every link connecting two nodes i and j has an associated weight w_{ij} which is a positive integer number.

The node degree k_i is defined as the number of edges incident upon a node. The in degree and out degree $k_i^{in/out}$ of node i is defined as the number of its in and out neighbours.

Degree centrality of the node i is the degree of that node. It can be normalised by dividing it by the maximum possible degree $N-1$:

$$dc_i = \frac{k_i}{N-1}. \tag{9.1}$$

Analogue, the in/out degree centralities are defined as in/out degree of a node :

$$dc_i^{(in/out)} = \frac{k_i^{(in/out)}}{N-1}. \tag{9.2}$$

Closeness centrality is defined as the inverse of farness, i.e. the sum of the shortest distances between a node and all the other nodes. Let d_{ij} be the shortest path between nodes i and j. The normalised closeness centrality of a node i is given by :

$$cc_i = \frac{N-1}{\sum_{i \neq j} d_{ij}}. \tag{9.3}$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Let σ_{jk} be the number of the shortest paths from node j to node k and let $\sigma_{jk}(i)$ be the number of those paths that pass through the node i. The normalised betweenness centrality of a node i is given by :

$$bc_i = \frac{\sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}}{(N-1)(N-2)}. \tag{9.4}$$

The strength of the node i is a sum of the weights of all the links incident with the node i:

$$s_i = \sum_j w_{ij}. \tag{9.5}$$

All given measures are defined for directed networks, but language networks are weighted, therefore, the weights should be considered. In the directed network, the in/out strength $s_i^{in/out}$ of the node i is defined as the number of its incoming and outgoing links, that is:

$$s_i^{in/out} = \sum_j w_{ji/ij}. \tag{9.6}$$

The selectivity measure is introduced in [19]. It is actually an average strength of a node. For the node i the selectivity is calculated as a fraction of the node weight and node degree :

$$e_i = \frac{s_i}{k_i}. \tag{9.7}$$

In the directed network, the in/out selectivity of the node i is defined as :

$$e_i^{in/out} = \frac{s_i^{in/out}}{k_i^{in/out}}. \tag{9.8}$$

9.5 Methodology

9.5.1 Data

For the network based keyword extraction we use the data set composed of Croatian news articles [107]. The data set contains 1020 news articles from the Croatian News Agency (HINA), with manually annotated keywords (key phrases) by human experts. The set is divided: 960 annotated documents for learning of supervised methods, and 60 documents for testing. The test set of 60 documents is annotated by 8 different experts, where the inter-annotator agreement in terms of F2 scores (see Section 5) are in average 46% (between 29.3% and 66.1%).

We selected the first 30 texts from the HINA collection for our experiment. The texts required some preprocessing: parsing only textual part and title part excluding annotations, cleaning of diacritics and symbols (w instead of vv, ! instead of l, etc.) and lemmatization. Non-standard word forms numbers, dates, acronyms, abbreviations etc. remain in text, since the method is preferably resistant to the noise presented in the data source.

The selected 30 texts varied in length: from very short 60 tokens up to 800 tokens (318 tokens in average). The number of keywords per document varies between 9 and 42 (24 in average). One annotator in average annotated 10 keywords per document.

9.5.2 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. Co-occurrence networks exploit simple neighbour relation, two words are linked if they are adjacent in the sentence [18]. The weight of the link is proportional to the overall co-occurrence frequencies of the corresponding word pairs within a corpus.

From the documents in the HINA data set we construct directed and weighted co-occurence networks: one from the text in each document and an integral one from the texts in all documents; 31 in total.

	TOP 10				TOP 24			
	$R[\%]$	$P[\%]$	$F1[\%]$	$F2[\%]$	$R[\%]$	$P[\%]$	$F1[\%]$	$F2[\%]$
In-degree	0	0	0	0	0.19	33.33	0.38	0.24
Out-degree	0	0	0	0	0.37	40.00	0.73	0.46
Closeness	0	0	0	0	0.75	**66.67**	1.48	0.93
Betweenness	0.19	**50.00**	0.38	0.24	0.37	50.00	0.73	0.46
In/out selectivity	**0.75**	40.00	**1.47**	**0.93**	**1.31**	29.17	**2.51**	**1.62**

Table 9.1: The top 10 and top 24 highly ranked keyword candidates form in-degree, out-degree, closeness, betweenness and in/out selectivity values obtained from all the HINA texts' network in terms of Recall (R), Precision (P), $F1$ and $F2$ score.

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 [8].

9.6 Keyword Extraction

9.6.1 Centrality Motivated Keyword Extraction

Network based keyword extraction methods exploit different measures for the task of identification and ranking the most representative features of the source - the keywords. The first part of our research compares the performance of different centrality motivated network measures (in/out degree, closeness and betweenness) with the performance of proposed selectivity measure. The second part develops a selectivity based method for keyword extraction with a comparative analysis of unsupervised (non-network enabled) approaches.

The degree (Eq. 9.1 and 9.2) of a node (word) is the number of neighbouring nodes (different neighbouring words). Typically, the nodes with the highest degree in the network are hubs, analogously the words with the highest degree are expectedly stopwords. The closeness (Eq. 9.3) of a node (word) is related to the farness of the word from all other words in the text. The betweenness (Eq. 9.4) of a node (word) is the measure of how many shortest paths between all other node-pairs are traversing a node. The words with the highest values of the betweenness centrality are considered to be important for the information flow as well. Selectivity is a local (node level) network measure, defined as the ratio of the node strength and the node degree. In weighted and directed co-occurrence networks one can consider the in- and out- links for obtaining in/out selectivity of the node (Eq. 9.8). The computation of the node's selectivity value is less complex and expensive than the computation of closeness and betweenness values.

From the network constructed from all the texts in the HINA news data set we calculate in/out degree, closeness, betweenness and in/out selectivity. Based on the obtained values we rank the top 10 or the top 24 keyword candidates from the network and evaluate them on the set of manually annotated keywords, as presented in Table 9.1. The top 10 or the top 24 keywords are selected due to the average number of human assigned keywords: in average 10 keywords from one annotator and in average 24 keywords from all 8 annotators per document. We evaluate the performance of each network measure based on standard recall (R), precision (P) and $F1$ score. $F1$ score is a harmonic mean of precision and recall: $F_1 = 2PR/(P+R)$. Beside the standard $F1$ score we also calculate the $F2$ score, which gives twice as much importance to the recall as to the precision: $F_2 = 5PR/(4P+R)$.

The results in Table 9.1 are in favour of the selectivity over other standard centrality network measures. The selectivity can efficiently differentiate between two basic types of nodes (words).

The nodes with high strength and high degree values, have low selectivity and they are usually closed-class words (e.g. stopwords, conjunctions, prepositions). The nodes with high strength and low degree have high selectivity values. Typically, the highest selectivity value nodes are open-class words which are preferred keyword candidates (nouns, adjectives, verbs) or even part of collocations, keyphrases, names, etc. On the other hand, the highest ranked words with in/out degree, closeness and betweenness are stopwords, which are not suitable keyword candidates. For example the top 10 ranked words according to in-degree centrality are: *to be, and, in, on, which, for, but, this, self, of*; according to betweenness they are: *to be, and, in, on, self, this, which, for, Croatian, but*; according to in/out selectivity they are: *Bratislava, area, Tuesday, inland, revolution, verification, decade, Balkan, freedom, Universe*.

In short, it seems that selectivity is insensitive to stopwords (the most frequent function words, which do not carry strong semantic properties, but are needed for the syntax of language) and therefore can efficiently detect semantically rich open-class words from the network and extract better keyword candidates.

9.6.2 Selectivity Based Keyword Extraction

The second part of our research develops a selectivity based method for keyword extraction. In order to compare the selectivity based extraction to non-network based approaches (unsupervised machine learning methods) we construct 30 networks (directed and weighted) from the 30 texts in the HINA data set and evaluate with manually annotated keyword sets .

From 30 networks we compute in/out selectivity for all nodes. The nodes are ranked according to the highest in/out selectivity values above a threshold value. Preserving the same threshold value (≥ 1) in all documents resulted in different number of nodes (one word long keyword candidates) extracted per each network. The obtained set of one word long keyword candidates is noted as SET1.

Then, for every filtered node we detect neighbouring nodes: for the in-selecti-vity we isolate one neighbour node with the highest outgoing weight; for the out-selectivity we isolate one neighbour node with the highest ingoing weight. The result of in/out selectivity extraction is a set of ranked word-tuples - SET2. Word-tuples are two-word long sequences of keyword candidates . From the obtained tuples we filtered out those containing stopwords in order to compare with the manually annotated evaluation set.

9.7 Evaluation and Results

For the keyword extraction task the strategy "more is better" can be utilized, since there is no objective judgement on keywords. Hence, it is preferable to extract more keywords which makes trade a off between precision and recall of the methods. The second polemic issue of keyword extraction task is: shorter keywords are more general vs. longer which are more accurate. Motivated by these open arguments, and by the approach of other authors, we decided to follow the same principle: to extract as many keyword candidates as possible and evaluate them on the basis of recall (R) and $F2$ score, beside the standard precision (P) and $F1$ score.

Evaluation is the final part of the experiment based on the intersection of the obtained sets SET1 and SET2 of keyword candidates with the union of all 8 annotators keywords. The results in terms of precision and recall are in Figures 9.1 and 9.2 respectively, and in terms of $F1$ and $F2$ scores in Figures 9.3 and 9.4 respectively. The obtained average $F1$ score for the SET 1 is 24.63%, and the average $F2$ score is 21.19%. The expansion of obtained candidates to SET2 increased the average $F1$ score to 25.9% and $F2$ score to 24.47%.

All supervised and unsupervised methods reported on keyphrases extraction from the HINA data set incorporate the linguistic knowledge (POS, MSD,..) of Croatian. Mijić et al. [107] initially

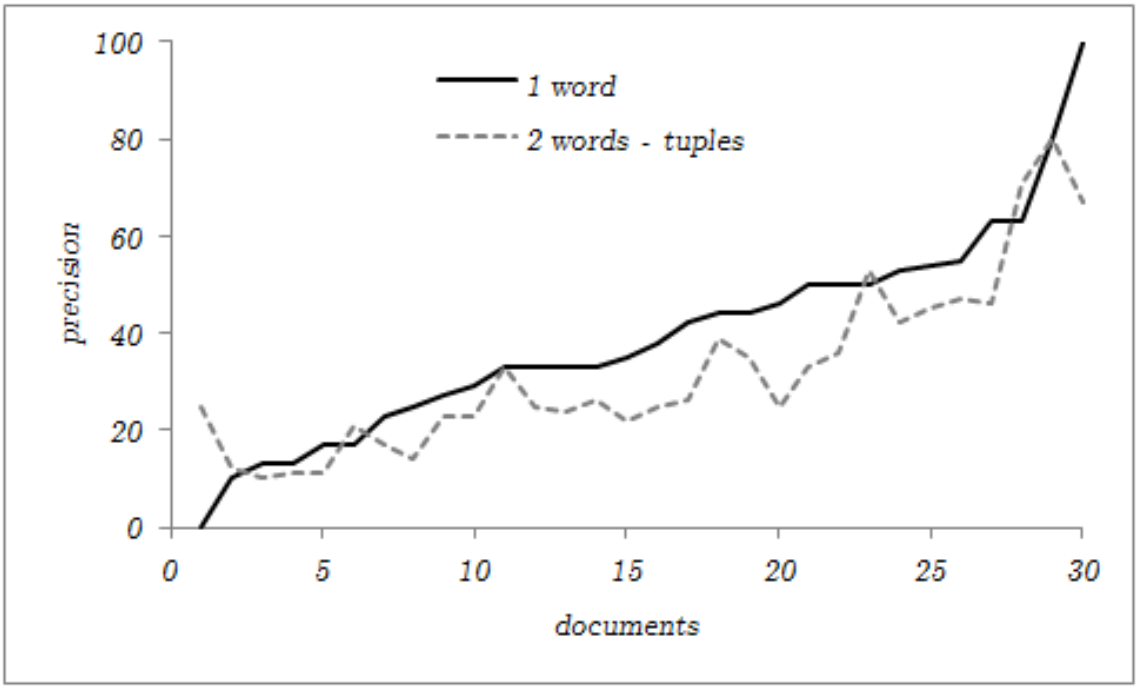


Figure 9.1: Precision on the SET1 (1 word candidates) and SET2 (2 word-tuples candidates) per 30 documents.

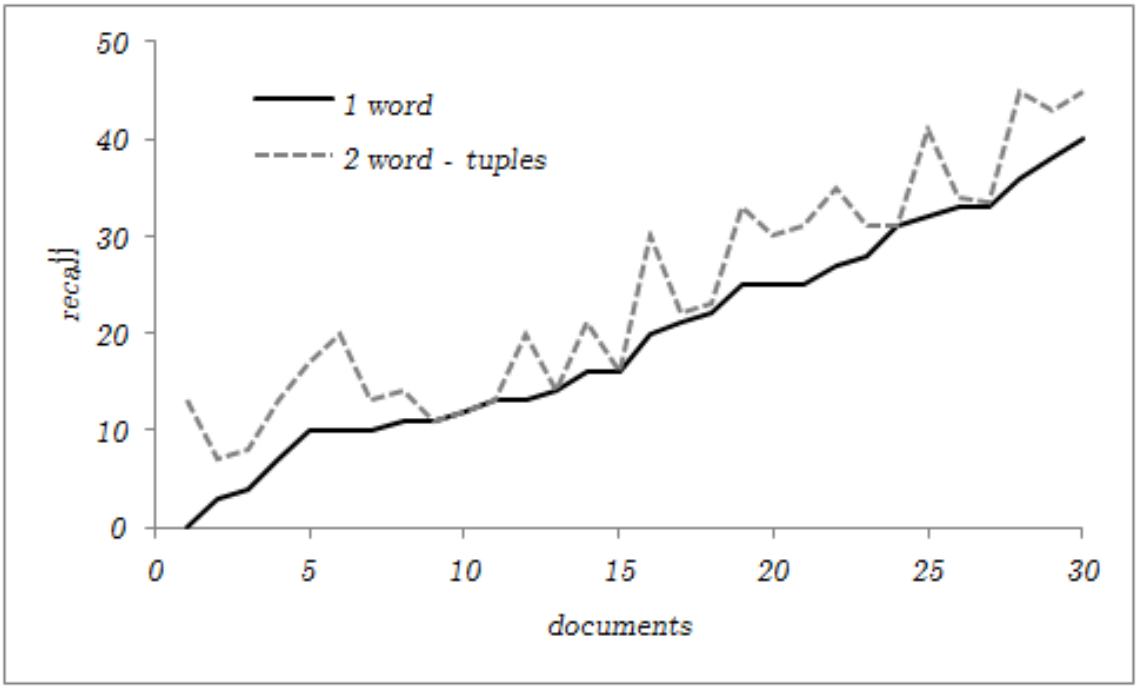


Figure 9.2: Recall on the SET1 (1 word candidates) and SET2 (2 word-tuples candidates) per 30 documents.

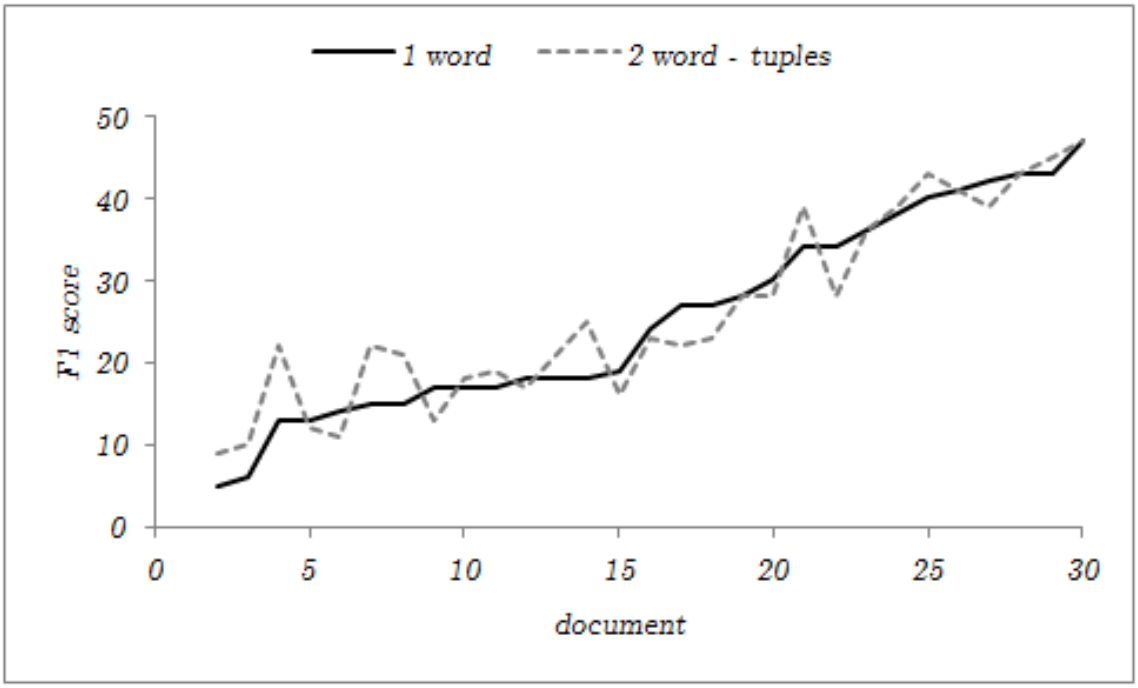


Figure 9.3: F1 score of the SET1 (1 word candidates) and SET2 (2 word-tuples candidates) per 30 documents.

extracted the list of keyword candidates as a comprehensive list of all words without stopwords) which was expanded into longer n-gram sequences up to a length of four. In [107] a keyphrase extraction system developed for a large-scale Croatian news production system the tf-idf ranking model was used to extract n-grams of up to length of four, which were lemmatized, and POS and MSD filtered. For evaluation the manually annotated key phrases from 60 documents were used. The evaluation set was reduced to keywords suggested only by 3 top annotators (having the highest inter-annotator agreement among all 8 annotators). The results indicate that the performance is comparable to that of the human annotators. Ahel et al. [76] for the one-word long keywords reported precision of 22% and recall of 3.4%.

We designed our method purely from statistical and structural information encompassed in the source text which is reflected in the structure of the network. Our method achieved on a SET1 average recall of 19.53% and precision of 39.1%. Expansion to the word-tuples in SET2 increased average recall to 23.87% and decreased precision to 32.23%. The obtained results are comparable to [107] and [76], but with a slightly different evaluation set up.

The obtained selectivity based results are promising and have potential to improve in several directions which is elaborated at the end of the next Section. An additional remark regarding results, is that beside keyword candidates our method captures personal names and entities, which were not marked as keyphrases and lowered the score. Capturing names and entities can be of high relevance for the tasks such as name-entity recognition, text summarization, etc.

Keyword annotation is an extremely subjective task as even human experts have difficulties to agree upon keyphrases (inter-agreement around 40%). Croatian is a highly morphologically rich language, which puts another magnitude of challenge on the task, since annotators are freely choosing the morphological word form as a tag, which seems appropriate at the moment. Additionally, there was no predefined set of index or keywords list, so annotators could make up their own, even worse in some cases it seemed appropriate to annotate with keywords, which were not present in the original article (out-of-vocabulray words). In [76] the number of out-of-vocabulary keywords

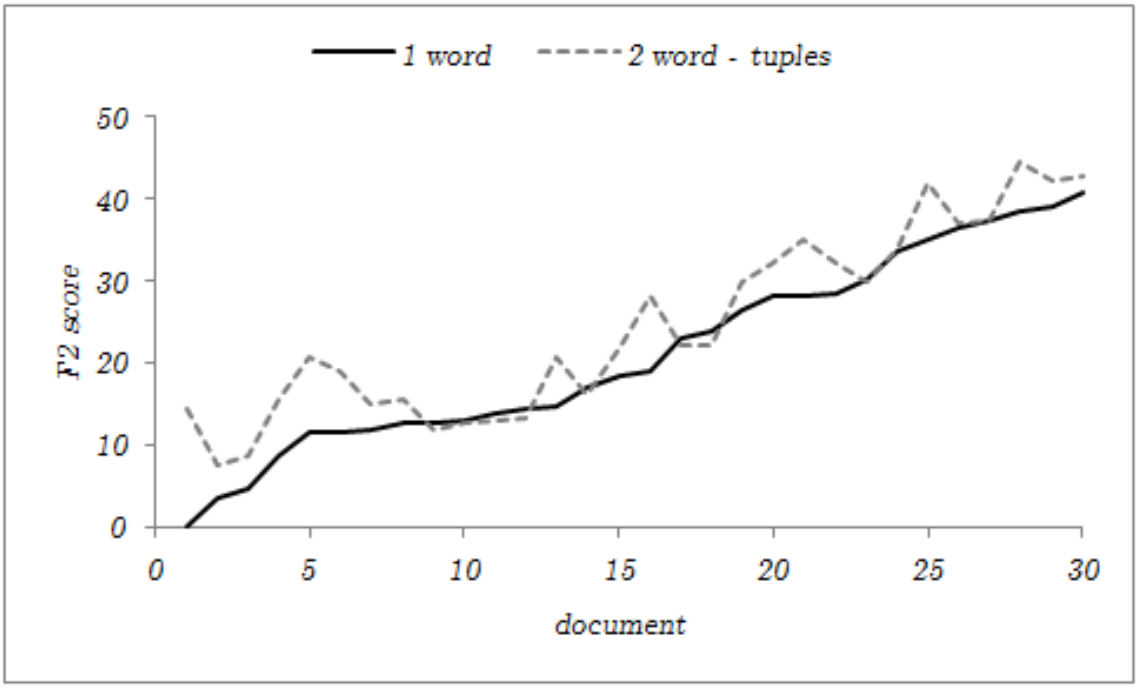


Figure 9.4: F2 score of the SET1 (1 word candidates) and SET2 (2 word-tuples candidates) per 30 documents.

on the whole of the HINA data set is estimated to a high of 57%. Since our method is derived from purely text statistics, it is not capable to capture all the possible subjective variations of the annotators or out-of-vocabulary words. Still it is close to the range of the inter-annotator achieved agreement .

9.8 Conclusion

This research on selectivity based keyword extraction for Croatian news (HINA data set) describes an unsupervised method which extracts nodes from a complex network as keyword candidates. We build our approach with a new network measure - the node selectivity (defined as the average weight distribution on the links of the single node). The node selectivity value is used for extracting and ranking the keyword candidates. Initially, we compare selectivity extraction to standard centrality motivated measures, and propose the selectivity measure for the keyword extraction.

The selectivity based keyword extraction method is comprised of: the extraction of the seed keyword set (words with the highest in/out selectivity) and expanding them to word-tuples with the highest in/out selectivity values. The obtained average $F1$ score for the set of extracted keyword candidates is 24.63%, and the average $F2$ score is 21.19%. The expansion of the obtained candidates to word-tuples increased the average $F1$ score to 25.9% and $F2$ score to 24.47%, which is comparable to the results on the same data set achieved by supervised and unsupervised methods, and is close to the range of the inter-annotator achieved agreement. The selectivity based extraction does not require linguistic knowledge as it is purely derived from statistical and structural information encompassed in the source text which is reflected in the structure of the network.

Our results imply that the structure of the network can be applied to the Croatian keyword extraction task with many possible improvements. This should be thoroughly examined in future work, which will cover: a) evaluation - considering all flective word forms; considering various matching strategies - exact, fuzzy, part-of-match; b) text types - considering texts of varying length,

genres and topics; c) multitopic - comparing isolate document extraction vs. multitopic extraction; d) other languages - testing on standard English (and other) data sets; e) longer keyword candidate sets - constructing keyword sequences up to a length of 3; f) entity extraction - testing weather entities can be extracted from complex networks.

10. Comparison of the Language Networks from Literature and Blogs

10.1 Abstract

In this Chapter we present the comparison of the linguistic networks from literature and blog texts. The linguistic networks are constructed from texts as directed and weighted co-occurrence networks of words. Words are nodes and links are established between two nodes if they are directly co-occurring within the sentence. The comparison of the networks structure is performed at global level (network) in terms of: average node degree, average shortest path length, diameter, clustering coefficient, density and number of components. Furthermore, we perform analysis on the local level (node) by comparing the rank plots of in and out degree, strength and selectivity. The selectivity-based results point out that there are differences between the structure of the networks constructed from literature and blogs.

10.2 Introduction

The representation and analysis of written texts in terms of graphs and complex networks offers an alternative approach for studying the language with different applications in the domain of natural language processing (NLP). Various types of linguistic networks have already been studied: syntax networks [23, 24], semantic networks [3], phonological networks [25], syllable networks [30, 31], word co-occurrence networks [7, 9–11, 14, 18, 19, 21, 36–38, 138, 139]. In [3, 5, 22] a systematic methodological overview of linguistic complex networks principles is presented. Recently, linguistic co-occurrence networks have been intensively studied in order to analyse the structure of the language [7, 9–11, 14, 18, 19, 21, 36–38, 138, 139].

As the networks incorporate associations between words and concepts, their structure, quantified by global and local network measures [140], such as clustering coefficient, shortest path, diameter, density, node degree, can provide information on some properties of the text . The motivation of our research was to find which network measures are sensitive on different texts categories.

In our previous research [18, 21] we showed the advantages of using a directed and weighted co-occurrence network as the model to capture the structure of a text. In this work we study global

and local network measures for the networks constructed from different categories of texts. In particular, at the local level, we applied the node selectivity measure in order to examine if it is sensitive on different styles of texts. Node selectivity is defined as the average weight distribution on the links of the single node [19]. Therefore, in our approach we constructed directed and weighted co-occurrence networks from different texts: 4 books and 4 blogs. We compare global and local network measures for book-blog network pairs.

In the Section 10.3 we present the overview of related work. In the Section 10.4 we present key measures of complex networks. In the Section 10.5 the data and network construction techniques are presented. In the Section 10.6 we present the results. In the last Section 10.7 we elaborate on the obtained data and provide concluding remarks.

10.3 Related Work

Ferrer i Cancho and Solé in [9] first showed that the co-occurrence networks have 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 [7] used co-occurrence networks to study language as a self-organising network of interacting words. Masucci and Rodgers in [11] investigated the co-occurrence network topology of Orwell's '1984' focusing on the local properties: nearest neighbours and the clustering coefficient. Furthermore, in [19] they introduced the node selectivity measure that can distinguish the difference between normal and randomised text. Liu and Cong [10] constructed co-occurrence networks from text in different languages and used complex network parameters for the classification (hierarchical clustering) of 14 languages, where Croatian was amongst 12 Slavic.

Different applications of linguistic network analysis in NLP includes: evaluation of language complexity [138], automatic summarisation [37], evaluation of machine translation [139], authorship attribution [36] and text quality analysis [38]. Costa et al. [138] studied the relationship between the topology of network and complexity of the text. They studied texts with different levels of simplification in co-occurrence networks and found that topological regularity correlated negatively with textual complexity.

In [37] the authors describe a method that uses complex networks concepts for the summarisation task. In [139] several metrics from complex networks are exploited in order to evaluate the quality of translations. The best distinctions were obtained with the out-degree, in-degree, minimum path and cluster coefficient. In [36] authors investigate the correlation between the properties of networks and author characteristics. It is shown that the networks produced for each author are sensitive to specific features, which indicates that complex networks can capture author characteristics and, therefore, could be used for the authorship identification. In [38] authors investigate the possibility of automated evaluation of text quality using topological measurements extracted from the corresponding complex networks. All the measures are correlated with grades assigned by human experts. The results indicate that, the presented approach has a potential to be used in the process of text quality evaluation.

10.4 The Network Structure Analysis

This Section contains explanations of the most important measures for network analysis. Every network has an N number of nodes and K number of links. Considering the fact that our networks are weighted every link connecting two nodes has an associated weight. The degree of a node i is the number of links with which the node is connected, k_i. In the case of the directed network, there are two kinds of the degree: the in-degree, k_i^{in} corresponding to the number of incoming links and

the out-degree, k_i^{out} equal to the number of outgoing links. The average degree of the network is:

$$< k >= \frac{2K}{N}. \tag{10.1}$$

For every two connected nodes i and j the number of links lying on the shortest path between them is represented as d_{ij}, and so d_i is the average distance of a node i from all other nodes, and it's obtained by:

$$d_i = \frac{\sum_{i \neq j} d_{ij}}{N}. \tag{10.2}$$

For the next two measures, if a network contains more than one component, we consider the largest component. The average shortest path length between every two nodes in network is:

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}, \tag{10.3}$$

and the maximum distance results in the network diameter, D:

$$D = max_i d_i. \tag{10.4}$$

The clustering coefficient is a measure which defines the presence of connections between the nearest neighbours of a node. And so, c_i (clustering coefficient) of a node is a fraction between the number of edges E_i that exist between that k_i and the total possible number :

$$c_i = \frac{2E_i}{k_i(k_i-1)}. \tag{10.5}$$

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

$$C = \frac{1}{N} \sum_i c_i. \tag{10.6}$$

Density of network is a measure of network cohesion defined as the number of observed relationships divided by the number of possible relationships

$$d = \frac{K}{N(N-1)}. \tag{10.7}$$

Strength of the node i is the number of its outgoing and incoming links (sum of its weights). We define the in-strength and the out-strength:

$$s_i^{out/in} = \sum_j w_{ij/ji}. \tag{10.8}$$

Node selectivity is a measure introduced in [11] that can capture the effective distribution of numbers in the weighted adjacency matrix, and it's obtained as a ratio of node strength and its degree :

$$e_i^{out/in} = \frac{s_i^{out/in}}{k_i^{out/in}}. \tag{10.9}$$

In order to illustrate the relationships between node degree, node strength and node selectivity, we constructed a small network of seven nodes presented in Figure 10.1. Additionally, Table 10.1 contains values of in-degree, out-degree, in-strength, out-strength and in-selectivity and out-selectivity for all seven nodes in the network presented in Figure 10.1.

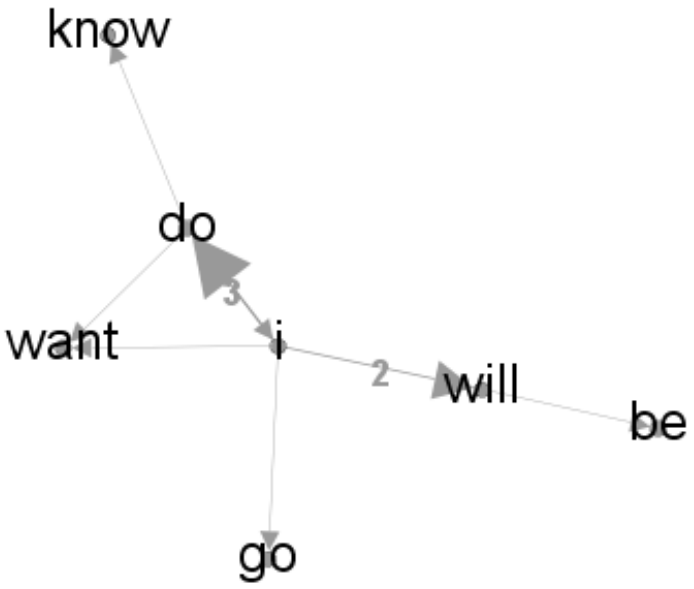


Figure 10.1: Weighted and directed co-occurrence network of seven nodes.

NODE	k_{in}	k_{out}	w_{in}	w_{out}	s_{in}	s_{out}
i	1	4	1	7	1	1,75
do	1	2	3	2	3	1
will	1	1	2	1	2	1
want	2	0	2	0	1	0
know	1	0	1	0	1	0
go	1	0	1	0	1	0
be	1	0	1	0	1	0

Table 10.1: Values of in/out - degree, strength and selectivity.

10.5 Network Construction

10.5.1 Data

Our corpus contains 4 books written or translated into the Croatian language, and 4 blog texts written in Croatian language. The books are: *Picture of Dorian Gray, Bones, The Return of Philip Latinowicz and Mama Leone*. The blogs are: *Index.hr, Slobodna Dalmacija, Narodne novine and Gospodarski list* (daily newspaper portal, or business portal). The feature which prompted us to do the comparison is the linguistic distinction between book and blog. Books are written in formal language, standard expressions and phrases are used, whilst blogs are mostly written in a casual mode, with the use of slang, the shortenings of the words or mistakes in syntax. Books come in different sizes and so we compared them with the approximately same sized blog (with the same number of different words), which means we had 4 book-blog pairs for comparison. The sizes of books and blogs in number of total words are shown in the first row of Table 10.2, while the numbers of different words are presented in the second row (as the number of nodes).

Measure	Text Bones	Gospodarski list	Mama Leone	Narodne novine
Number of different words	191 986	199 417	85 347	146 731
Number of nodes (N)	27396	27727	13067	13036
Number of edges (K)	102052	105171	49383	55661
Average degree ($< k >$)	7,45	7,58	7,56	8,54
Avg. shortest path (L)	3,21	3,28	3,11	3,17
Diameter (D)	10	21	10	12
Avg. clust. coeff. (C)	0,25	0,22	0,29	0,22
Density (d)	0,0002	-	0,00056	0,00066
No. connect. compon.	15	7	1	2

Measure	Text Picture of Dorian Gray	Index.hr	Return of Phillip Latinowicz	Slobodna Dalmacija
Number of different words	75 099	118 548	28 137	44 367
Number of nodes (N)	15631	15065	9531	9553
Number of edges (K)	46201	28972	21760	25155
Average degree ($< k >$)	3,88	3,85	4,57	5,27
Avg. shortest path (L)	3,45	3,45	3,59	3,56
Diameter (D)	14	22	16	13
Avg. clust. coeff. (C)	0,18	0,016	0,15	0,17
Density (d)	0,0004	0,0002	0,00048	0,00055
No. connect. compon.	1	45	5	3

Table 10.2: The comparison of network measures for book-blog network pairs.

10.5.2 The Construction of Co-occurrence Networks

We used Python and the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [8]. The first step in creating networks was text 'cleaning': normalising symbols for Croatian diacritics (č, ć, ž, đ, and š), removing special symbols and removing punctuation which does not mark the end of a sentence. We created 8 networks, weighted and directed. Nodes are words that are linked if they are direct neighbours in a sentence. The next step was creating the networks as weighted edgelists , which contain all the pairs of connected words and their weights (the number of connections between two same words).

10.6 Results

In this Section we present the results of our measuring described in Section 2, such as average degree $< k >$, average path distance L, diameter D, and the average clustering coefficient C, density d, node strength s_i and node selectivity e_i. In Table 10.2 we present the estimated global network measures. There are certain differences between measures for book-blog network pairs, but there is no uniform rule that may be used to differentiate between these two styles of writing.

Furthermore, we compare networks on the node-level using degree, strength and selectivity measures. For the purpose of comparison we used rank plots. The in/out-degree rank function represents the relationship function between the rank and the in/out-degree of the degree sequence of all nodes sorted in decreasing order. Similarly, the in/out-strength rank plot and the in/out-selectivity

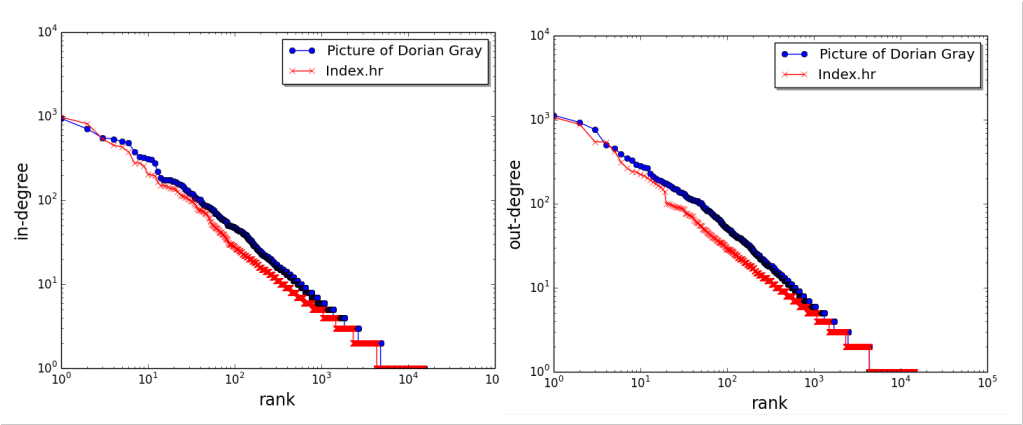


Figure 10.2: In-degree and out-degree rank plots.

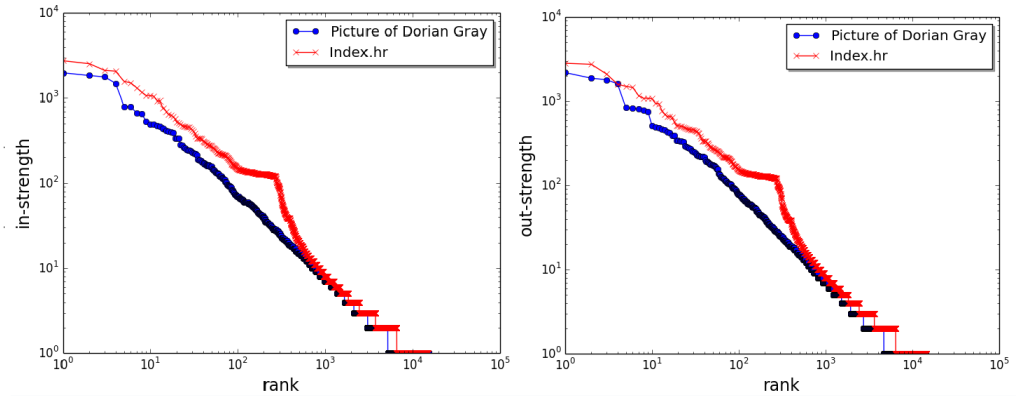


Figure 10.3: In-strength and out-strength rank plots.

rank plot are defined.

The results of the comparisons of the in-degree rank plot and the out-degree rank plot for one book-blog network pair are shown in Figure 10.2. The plots do not show significant difference for the in-degree nor for the out-degree rank plots between book-blog network pair. We also experimented with additional three book-blog pairs and we obtained similar results (not reported here due to limited space).

The results of the comparisons of the in/out-strength rank plot for the same book-blog network pair are shown in Figure 10.3. Again, there is no difference. Except some small deviation that can be noticed in the plot, but we cannot conclude that in/out-strength distinguish books from blogs.

The selectivity rank plots are shown in Figure 10.4 (in-selectivity) and in Figure 10.5 (out-selectivity). The results show that there are differences in selectivity rank plots between networks constructed from books and networks constructed from blogs for all 4 book-blog network pairs. In general, all in/out-selectivity values are lower for books than for blogs. We disregarded nodes with zero values of degree because it causes the division by zero (in total 4% of nodes).

10.7 Conclusion

In this work we analysed which complex network measure can distinguish between different text categories: literature and blogs. Our results indicate that global network measures are not precise enough to capture the structural differences between networks constructed from different text

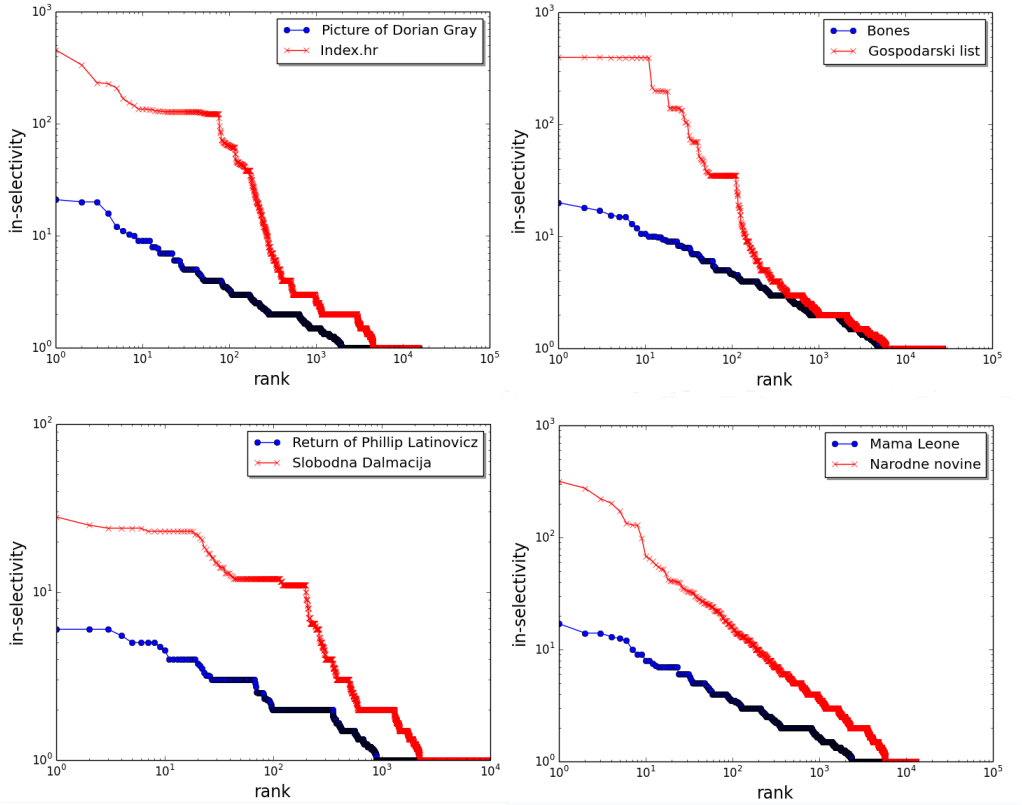


Figure 10.4: In-selectivity rank plots for 4 book-blog network pairs.

categories. Even the compared in/out- degree rank plots and in/out- strength rank plots do not clearly show the differences. However, in-selectivity and out-selectivity rank plots indicate that there are structural differences between networks constructed from books and networks constructed from blogs.

Similar approaches of complex network based analysis have been used in certain applications in the domain of NLP. In [138] it is shown that strength, shortest path, diversity and hierarchical measures can make a distinction between normal text and simplified text. In [138] it is presented how in-degree, out-degree, minimal path and clustering coefficient can be used for machine translation evaluation. In [138] it is shown that out-degree, clustering coefficient and deviation from linear dynamics in the network growth are correlated with the text quality. However, there are no comprehensive studies focused on finding network measures that are sensitive to different text categories. Our work is the first attempt to analyse whether the node selectivity measure can differentiate between books and blogs networks. These results can be further tested on various categories of texts.

For future work we will examine other local network measures that depend on the strength, degree and link direction in combination with other measures such as clustering coefficient. Finally, these results encourage us to investigate the complex network properties for text classification, text evaluation or even text quality assessment.

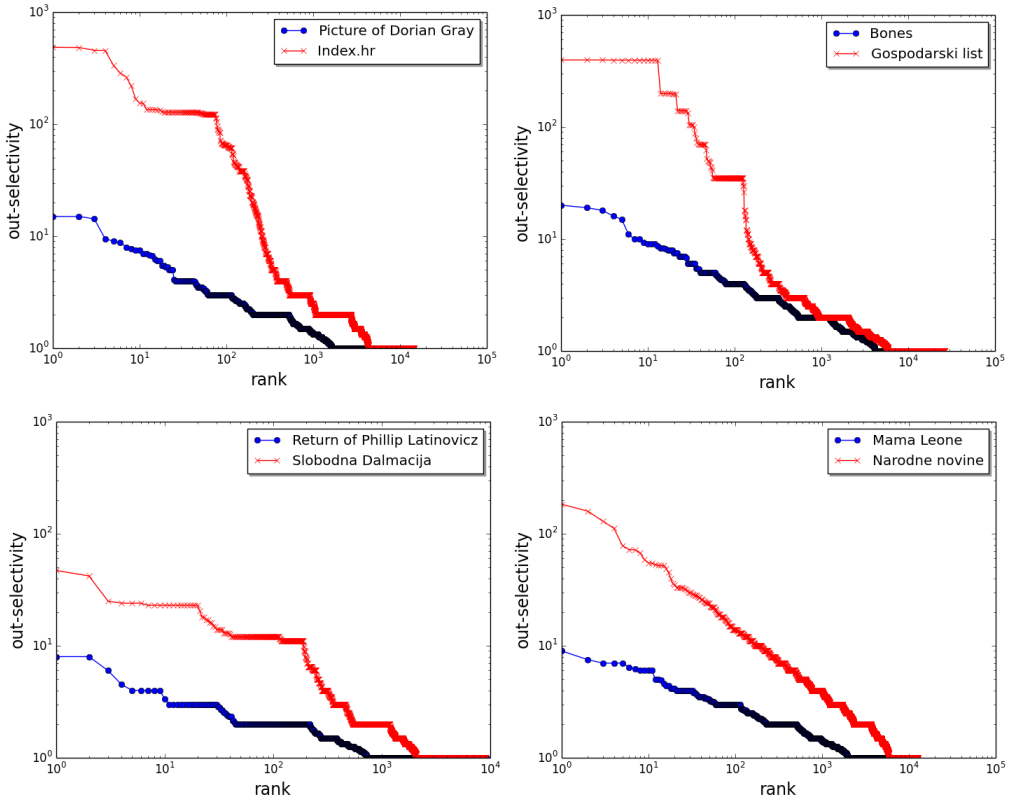


Figure 10.5: Out-selectivity rank plots for 4 book-blog network pairs.

11. Revealing the Structure of Domain Specific Tweets via Complex Networks Analysis

11.1 Abstract

In this work we explore the relation between different groups of tweets using complex network analysis and link prediction. The tweets were collected via the Twitter API depending on their textual content . That is, we searched for the tweets in English language containing specific predefined keywords from different domains. From the gathered tweets a complex network of words was formed as a weighted network. Nodes represent words and a link between two nodes exists if these two words co-occur in the same tweet, while weight denotes the co-occurrence frequency. The Twitter search was repeated for four different search criteria (API queries based on different tweet keywords), thus resulting in four networks with different nodes and links. The resulting networks were subjects to further network analysis, as comparison of numerical properties for different networks and link prediction for individual networks. This work shows the tweet scraping process, our approach to building the networks, the measures we calculated for them, the differences and similarities between different networks we built and our success in predicting future links.

11.2 Introduction

Twitter is a popular online social network created in 2006 that enables user to send publicly visible messages called "tweets" . One of the main characteristics that distinguishes Twitter from other online social networks is the limit on tweet length. Twitter user are allowed to send tweets that have a maximum of 140 characters. Hence, Twitter is often categorized as a micro-blogging platform . It is estimated that in 2015 Twitter had over half a billion users. [141]

Because of its popularity, user-base size and vast amounts of tweets, Twitter has been studied in the context of person-to-person relations [142], user influence [143], economic predictions [144], predictions of political elections [145], conversational practices [146] and trends discovery [147] .

Another important research domain related to Twitter is sentiment analysis. In [149] Pak et al. automatically collect from Twitter a corpus and perform linguistic analysis on it. Then they build a sentiment classifier able to determine positive, negative and neutral sentiments for a

document. There has been reported research in automatic classification of tweets regarding their sentiment [148]. [150] gives a detailed revision of the field of sentiment analysis with Twitter in focus. Research by Agarwal et al. [151] examines sentiment analysis on Twitter data. In it the authors introduce POS-specific prior polarity features and explore the use of a tree kernel to eliminate the need for laborious feature engineering. In [152] Kouloumpis et al. investigate the utility of linguistic features for detecting tweets sentiment using a supervised approach, while also leveraging existing hashtags in building training data . Wang et al. [153] present hashtag-level sentiment classification which aims to automatically generate the overall sentiment polarity for a given hashtag in a certain time period.

The following papers use the complex network analysis approach to Twitter data. Villazon et al. in [154] look at Twitter as a complex network, calculating the cluster coefficient, power law and average path length for it. [155] presents a model for describing the growth of scale-free networks. The model is applied only after checking that Twitter is indeed a scale-free network, and for that purpose the mentioned paper proposes a new heuristic method of finding the upper bounds of the path lengths instead of computing the exact length .

In our approach we use complex networks analysis to reveal the structure of domain specific tweets. The motivation of our research is to detect weather networks constructed from different tweets domains have different structural properties. More precisely, the goal of this research is to determine whether (and which) complex network measures can distinguish between networks of tweets with "positive" and "negative" aspects. Possible applications of proposed approach can be in the domain of sentiment analysis. Furthermore, link prediction enables anticipation of positive or negative attitude propagation on Twitter.

We collect positive tweets in English language using keywords with positive polarity (e.g. joy, happiness, ...) and negative tweets using keywords with negative polarity (e.g. anger, fear, ...). Then we perform the global and local complex network analysis where we compare results for four obtained networks. On the global level we use a standard set of network measures (e.g. diameter, average path length, clustering coefficient). However, for the local level analysis we apply a node selectivity measure encouraged by our previous findings [50, 59, 71] for which we show that it is an important measure for language networks analysis and differentiation.

In the Section 11.3. we present the network measures used in our research. In the Section 11.4. we describe how we construct the tweet networks. The results and discussion are given in the Section 11.5. Finally, the Section 11.6 contains conclusions and directions for the further research.

11.3 Networks Measures

Complex network is a graph with non-trivial topological features (e.g. high clustering coefficient, low distances, heavy-tailed degree distribution, etc.). It can be represented with a graph G, defined as a pair of two sets $G = (V, E)$; the first set V consisting of vertices and the second set E consisting of edges. N as the number of vertices in V and K as the number of edges in E. In the domain of network analysis, the vertices are referred as nodes and the edges are called links.

Network analysis can be classified by the following three levels: macro-scale or global level, meso-scale level and micro-scale or local level. In weighted complex networks every link connecting two nodes u and v has an associated weight w_{uv}. A node degree is the number of links directly connected (or incident) to that node. The set of nodes incident to a node v is denoted as $\Gamma(v)$. The number of network components is represented by ω. Next, we present network measures that will be used in the following sections.

The average network degree is the ratio of the number of links to the number of nodes. For undirected networks we multiply this ratio by 2 since undirected links always have two incident

nodes:

$$\langle k \rangle = 2\frac{K}{N}. \tag{11.1}$$

Network strength is simply the sum of all link weights in a network:

$$S = \sum_{u,v \in V} w_{uv}. \tag{11.2}$$

For the average network strength we divide a networks strength with its number of nodes:

$$\langle s \rangle = \frac{S}{N}. \tag{11.3}$$

Node selectivity for a node v corresponds to the sum of weights of all incident links divided by that nodes degree (denoted as $\deg(v)$):

$$e(v) = \frac{\sum_{u \in \Gamma(v)} w_{uv}}{\deg(v)}. \tag{11.4}$$

Average network selectivity is the sum of all individual node selectivities divided by the number of nodes:

$$\langle e \rangle = \frac{\sum_{v \in V} e(v)}{N}. \tag{11.5}$$

Network density is represented as the ratio between the number of existing links and the number of all possible links:

$$d = \frac{K}{N(N-1)}. \tag{11.6}$$

Average path length for a network, where d_{uv} denotes the number of links lying on the shortest path between $u, v \in V$, is computed as following:

$$L = \sum_{u,v} \frac{d_{uv}}{N(N-1)}. \tag{11.7}$$

The network diameter represents the longest shortest path in a network ($u, v \in V$):

$$D = \max(d_{uv}). \tag{11.8}$$

The network radius denotes the shortest $\varepsilon(v)$, where $\varepsilon(v)$ is defined as the maximum distance between $v \in V$ and any other node:

$$R = \min(\varepsilon(v)). \tag{11.9}$$

Network transitivity where possible triangles are identified by the number of triads (two links with a shared node) :

$$T = 3\frac{\#triangles}{\#triads}. \tag{11.10}$$

Average clustering coefficient, where $c(v)$ is the clustering coefficient for a node v, sums all the individual clustering coefficients and divides them by the number of nodes:

$$C = \frac{1}{N}\sum_{v\in V} c(v). \tag{11.11}$$

The global network efficiency is the reciprocal value of a networks average path length :

$$E = \frac{1}{L}. \tag{11.12}$$

In the context of link prediction we use the following measures.

Weighted Common Neighbors, adapted from [156], where weights of links connecting u and v to their common neighbors are summed :

$$CN(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} w_{uz} + w_{vz}. \tag{11.13}$$

Weighted Jaccard's Coefficient, adapted from [157], which divides the weighted Common Neighbors value for u and v by the summed weights of all links incident to u and/or v :

$$JC(u,v) = \frac{\sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} w_{uz} + w_{vz}}{\sum_{a\in\Gamma(u)} w_{au} + \sum_{b\in\Gamma(v)} w_{bv}}. \tag{11.14}$$

Lastly, we present the link prediction precision as the ratio between the number of correctly predicted links and the total number of predicted links. That is, we divide the number of true positives ($|TP|$) by the number of true and false positives ($|TP| + |FP|$). [158]

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \tag{11.15}$$

11.4 Networks Construction

The first step in constructing networks is the collection of data. Initially, we searched for four sets of tweets according to the following criteria: a) tweets associated to recent immigrant and war related events; b) tweets containing negatively polarized words; c) tweets associated to house pets and d) tweets containing positively polarized words. The subset of positive and negative polarized words is extracted from the sentiment lexicon in [159]. From now on we will refer to the networks built from their respective sets as: a) emo-net^a, b) emo-net^b, c) emo-net^c and d) emo-net^d.

For the data collection process we use Python in combination with the Python Twitter Tools package, which provides an easy-to-use interface for the official Twitter API. In the API request arguments we specified we are searching for a mix of recent and popular tweets in the English language. We scraped about 10000 tweets for each of four different queries, resulting in a dataset

of 39882 tweets. It is worth to mention that the official Twitter API documentation states that the language detection is based on the "best-effort" principle [160].

In the text (tweets) preparation step first we eliminate stopwords[1], and from the remaining text we compute the 100 most frequent words for each of the four subsets. We selected top 100 words as the reasonable list which provides the best trade-off between computation time and link prediction results. Note that the former computation was case-insensitive and we used the list of English stopwords presented at `http://www.ranks.nl/stopwords`.

From the words of preprocessed tweets extended with the set of explicit keywords (e.g. joy, puppy) used for retrieving each of the tweets we form the nodes of the networks. Link between two nodes (words) is established if these two word appear together in the same tweet. Weight on the link represents words co-occurrence frequencies, that is, the number of tweets in which two high-frequency words from the top 100 list co-occurred. That makes the generated networks weighted and undirected . Hence, based on the high-frequency words, we generate four different networks for each of the four data sets.

We build 16 distinct networks from four datasets: the first network is built from 25% of the data, the second from 50%, the third from 75% and the fourth from 100% of the data in one dataset. We will denote those networks, respectively, as emo-net$_1^x$, emo-net$_2^x$, emo-net$_3^x$ and emo-net$_4^x$, where $x \in \{a,b,c,d\}$. That means we, as previously mentioned, generate a total of 16 different networks, four per each dataset.

Some other used Python packages not previously mentioned are NetworkX [8] and LaNCoA [161]. The first one is a popular Python tool for creating and manipulating complex networks. It also provides a rich collection of functions for studying complex networks on various levels. The LaNCoA toolkit provides procedures for construction and analysis of complex language networks .

11.5 Results

Global and local network measures

Here we present the computed global and local network measures for emo-net$_4^a$, emo-net$_4^b$, emo-net$_4^c$ and emo-net$_4^d$. Table 11.1 shows the calculated measures that were previously described in Section 11.3.

Measure	emo-net$_4^a$	emo-net$_4^b$	emo-net$_4^c$	emo-net$_4^d$
N	101	101	103	104
K	3454	3958	2854	3848
$\langle k \rangle$	68.396	78.3762	55.4175	74
$\langle s \rangle$	1025.9406	830.505	747.0291	1310.25
$\langle e \rangle$	29.4867	24.0104	42.9054	44.7693
d	0.684	0.7838	0.5433	0.7184
ω	1	1	1	1
L	1.316	1.2162	1.4582	1.2816
D	2	2	3	2
R	1	1	2	1
T	0.7965	0.875	0.7774	0.8595
C	0.0088	0.0208	0.0532	0.0077
A	-0.1257	-0.0933	-0.0587	0.0442
E	0.7599	0.8222	0.6858	0.7803

Table 11.1: Global and local network measures.

[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 a language (pronouns, prepositions, conjunctions, abbreviations, ...).

The first visualization we present (Figure 11.1) is for the node degrees across all emo-net$_4$ networks. We see no major differences for node degrees across those networks.

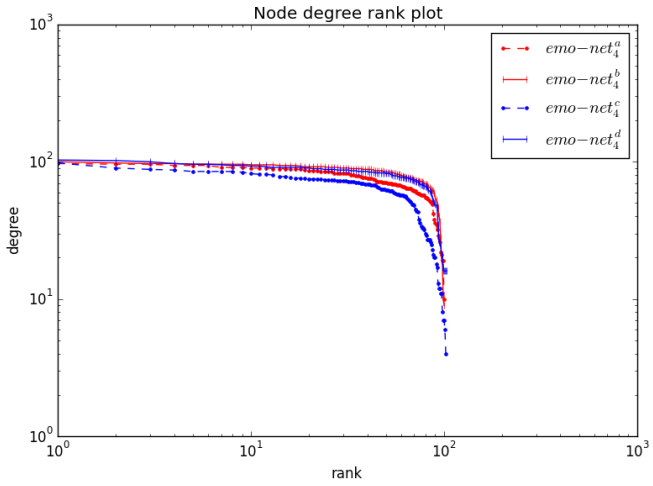


Figure 11.1: Node degrees for all emo-net$_4$ networks on a log-log scale.

Lets recall that emo-net$_4^a$ and emo-net$_4^b$ were based on data from queries with negative connotations. In contrast, emo-net$_4^c$ and emo-net$_4^d$ were based on queries with positive connotations. The most obvious difference between the first two "positive" and the last two "negative" networks in Table 11.1 is $\langle e \rangle$, which represent the value of average network selectivity. $\langle e \rangle$ is notably lower for emo-net$_4^a$ and emo-net$_4^b$ than for emo-net$_4^c$ and emo-net$_4^d$. Average network selectivity can be interpreted as how "heavy" the links across a network are. We see how our positive networks have on average stronger ties between nodes.

In Figure 11.2 we visualize the node selectivities for the networks mentioned above. Note that the plot in Figure 11.2 uses a log-log scale.

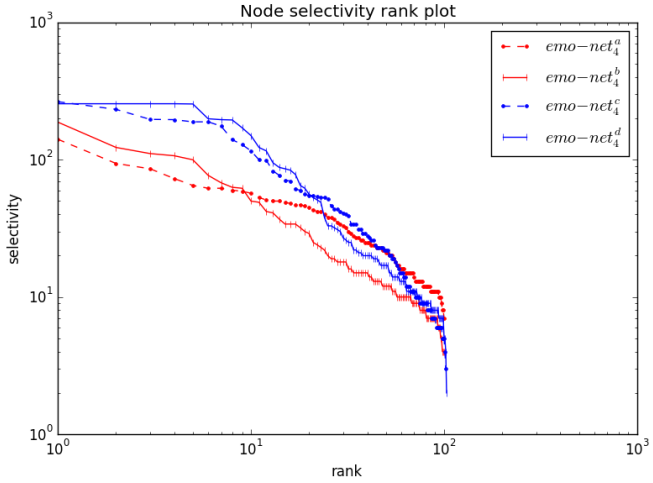


Figure 11.2: Node selectivities for all emo-net$_4$ networks on a log-log scale.

Link prediction

Next we present the results for the link predictions. Here we computed the most likely future links for emo-net$_1^x$, emo-net$_2^x$ and emo-net$_3^x$ where $x \in \{a, b, c, d\}$. The prediction were made using

two measures: weighted Common Neighbors (Table 11.2) and weighted Jaccard's Coefficient (Table 11.3). The definitions of both measures can be found in Section 11.3.

We will briefly describe the link prediction process which is the same for both measures. First compute the ranks for all non-existing links in emo-net$_i^x$, $x \in \{a,b,c,d\}$, $i \in \{1,2,3\}$. Generate the first set that contains the top n ranked non-existing links in emo-net$_i^x$ (n is the number of new links in emo-net$_{i+1}^x$). Next, generate the second set that holds links which appear in emo-net$_{i+1}^x$ but not in emo-net$_i^x$. Calculate the prediction precision by looking at the intersection of the first and second set.

Network	emo-net^a	emo-net^b	emo-net^c	emo-net^d
(25%) emo-net$_1$	29.96%	45.92%	30.13%	26.84%
(50%) emo-net$_2$	24.57%	34.88%	21.43%	17.73%
(75%) emo-net$_3$	28.49%	27.49%	18.4%	13.45%

Table 11.2: Prediction precision based on the weighted Common Neighbors measure.

Network	emo-net^a	emo-net^b	emo-net^c	emo-net^d
(25%) emo-net$_1$	35.88%	47.96%	37.95%	50.26%
(50%) emo-net$_2$	29.69%	37.72%	28.97%	41.14%
(75%) emo-net$_3$	12.85%	32.16%	30.06%	32.75%

Table 11.3: Prediction precision based on the weighted Jaccard Coefficient measure.

We see from Tables 11.2 and 11.3 that all predictions had a precision rate above 10%, with some going as high as 50%. The predictions are, by a large margin, most precise for emo-net$_1$ networks. Generally, those networks will not have all of the probable links already in them. With more data all the probable links are added. In most cases the prediction precision for networks with more links tends to fall. That is the only obvious trend for precision rates across query domains, network sizes and prediction measures.

11.6 Conclusion

In this Chapter we present how we construct multiple complex networks based on four different data sets. Each data set featured a collection of tweets gathered by predefined Twitter API queries. Two of those queries retrieved "negative" oriented tweets, while the other two gathered "positive" oriented tweets. We investigate global and local network measures across four query categories and compare them between "negative" and "positive" networks. In this Chapter we also predict future links for networks across all query domains. For that purpose we use networks built form a lower percentage of data and compare them with networks built from a higher percentage of the same data.

Regarding network measures, we found that the average network selectivity is the only measure that discriminates between "negative" and "positive" networks, favoring the positive ones. This preliminary results indicate that selectivity based network measures could be used in the Twitter sentiment analysis tasks.

The link prediction process gave no obvious patterns, except the higher prediction precision for networks built from the smallest amount of data. Also, for all of our networks the link prediction precision was above 10%. It should be noted that all our results are preliminary and a more complex analysis would be in order. Such analysis should primarily consider larger and more diverse data sets. Expanding the list of computed network measures would be also worth considering, along with community detection algorithms.

12. Link Prediction on Twitter

12.1 Abstract

With over 300 million active users, Twitter is among the largest online news and social networking services in existence today. Open access to information on Twitter makes it a valuable source of data for research on social interactions, sentiment analysis, content diffusion, link prediction, and the dynamics behind human collective behaviour in general. Here we use Twitter data to construct co-occurrence language networks based on hashtags and based on all the words in tweets, and we use these networks to study link prediction by means of different methods and evaluation metrics. In addition to using five known methods, we propose two effective weighted similarity measures, and we compare the obtained outcomes in dependence on the selected semantic context of topics on Twitter. We find that hashtag networks yield to a large degree equal results as all-word networks, thus supporting the claim that hashtags alone robustly capture the semantic context of tweets, and as such are useful and suitable for studying the content and categorization. We also introduce ranking diagrams as an efficient tool for the comparison of the performance of different link prediction algorithms across multiple datasets. Our research indicates that successful link prediction algorithms work well in correctly foretelling highly probable links even if the information about a network structure is incomplete, and they do so even if the semantic context is rationalized to hashtags.

12.2 Introduction

Our cumulative culture relies on our ability to carry the knowledge from previous generations forward. For millennia, we have been upholding a cumulative culture, which leads to an exponential increase in our cultural output [162], and it has given us evolutionary advantages that no other species on the planet can compete with. Unprecedented technological progress and scientific breakthroughs today make the amount of information to carry forward staggering. This requires information sharing, worldwide collaboration, the algorithmic prowess of search engines, as well as the selfless efforts of countless volunteers to maintain, categorize, and help navigate what we know. The task is made easier by the fact that much of what we know has been digitized [163, 164].

The combination of data deluge with recent advances in the theory and modeling of social systems and networks [42, 165–172] enables quantitative explorations of our culture that were unimaginable even a decade ago. Recent research has been devoted to enhanced disease surveillance [173], the spreading of misinformation [174, 175], to study human mobility patterns [176, 177] and the dynamics of online popularity [178], to quantify trading behavior [179, 180] and the dynamics of our economic life [181], as well as to study universality in voting behavior [182], political polarity [183] and emotional blogging [184, 185], to name just some examples.

The openness of Twitter to research has made it an important source of data for innovative data-driven research that lifts the veil on how we share information, how and with whom we communicate, and essentially on how we live our lives . Twitter was created in 2006, enabling users to send short publicly visible messages called tweets. Tweets typically consist of text, links (i.e. URLs), user mentions (with @ sign), retweet information (RT) and hashtags. Hashtags are marked with the # sign and are used for meta tagging, which enables users to find a specific theme or content [186] . Hashtags are neither limited nor do they have a predefined structure or content. Still they often capture the very essence of posted messages, much like keywords or keyphrases do [187], and they can be used effectively to monitor trends of topics on Twitter [147] as well as the polarity of tweets [153] . So far, Twitter data has been used to study the growth mechanisms of social interactions [155], for assessing user influence [143], for recommending (predicting) whom to follow [188], for information propagation [189], as well as for sentiment analysis [145, 150, 153].

Here we use Twitter data to study link prediction in the realm of co-occurrence language networks based on hashtags and based on all the words in tweets . Link prediction refers to inferring the future relationships from nodes in the complex network, or more formally, to estimate the likelihood of the existence of a link between two nodes based on the observed network structure and node attributes. A comprehensive review of link prediction methods is provided in [190]. In addition to relying on topological properties of networks, the problem was also addressed by the means of various machine learning techniques [157, 191]. Typical networks addressed by means of link prediction methods include protein-protein interaction networks and social networks, where one can predict longitudinal changes over time [190, 192–195]. While local similarity measures have traditionally been explored for unweighted networks, recently weighted local similarity measures have attracted more attention [156, 157, 196–198] . In line with these trends, we therefore focus on weighted local similarity measures for the prediction of links in the networks constructed from the content of tweets.

In addition to using five known methods, namely the weighted common neighbors (CN), the weighted Jaccard coefficient (JC), the weighted preferential attachment (PA), the weighted Adamic-Adar (AA) and the weighted resource allocation index (RA) [156, 157, 199], we also propose selectivity (SE) [11] and inverse selectivity (IS) as two effective weighted similarity measures. Selectivity is defined as the average weight distributed on the links incident to the single node, and has proven efficient for different language network tasks, ranging from the differentiation between original and shuffled text [21] to the differentiation of text genres [200] and for keyword extraction [201, 202]. We also note that link prediction on Twitter has been studied before in [203], where CN, AA, JC and RA measures were combined with the information about corresponding communities as determined with a variant of the label propagation algorithm in unweighted and directed networks. It was shown that this leads to an improvement of the area under the receiver operating characteristic curve (AUC) when structural measures are accompanied with community information to train supervised data mining models for link prediction. In [194] an approach has been proposed to predict future links in Twitter reciprocal reply networks by applying the covariance matrix adaptation evolution strategy to optimize weights based on neighbourhood and node similarity indices. It was shown that this method is suitable for predicting future followers on social networks.

As we will show after describing the Methods, our research reveals that hashtag networks yield to a large degree equal results as all-word networks, therefore supporting the claim that hashtags alone robustly capture the semantic context of tweets, and as such are useful and suitable for studying the structure of tweets. We will also show how introducing ranking diagrams is an efficient tool for the comparison of the performance of different link prediction algorithms across multiple datasets.

12.3 Methods

The network $G = (V,E)$ is a pair of a set of nodes V (or vertices) and a set of links E (or edges), where N is the number of nodes and K is the number of links. In weighted networks every link connecting two nodes u and v has an associated weight w_{uv}. A node degree $deg(u)$ is the number of links incident to node u and the set of neighbor nodes to a node u is denoted as $\Gamma(u)$. The strength of a node s_u is the sum of weights of all the links incident to u. More details about complex networks analysis can be found in [67] and all measures used for the quantification of the studied networks properties are listed in S1 Text.

There are various approaches for the link prediction task based upon similarity measures [190, 193]. In general each pair of nodes u and v ($u, v \in V$) is assigned a score p_{uv} which is directly defined as the similarity between nodes u and v. Then the link prediction task is to determine whether the link between u and v will be established according to the descending order of assigned scores p_{uv}. Next we define seven link prediction measures used in this study.

In the weighted common neighbors (CN) link prediction measure weights of links connecting nodes u and v to their common neighbors z are calculated as in [156]:

$$CN(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} (w_{uz} + w_{vz}) \tag{12.1}$$

where $\Gamma(u)$ and $\Gamma(v)$ are the sets of neighbors of nodes u and v. CN measures the number of neighbors that two nodes have in common, while for the weighted CN the sum of weights is used instead. CN is the simplest but at the same time computationally undemanding measure which serves as a baseline for link prediction.

The weighted Jaccard coefficient (JC) adapted from [157], divides the weighted common neighbors value for u and v by the sum of weights on all the links incident to u and/or v :

$$JC(u,v) = \frac{\sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} (w_{uz} + w_{vz})}{\sum_{a\in\Gamma(u)} w_{au} + \sum_{b\in\Gamma(v)} w_{bv}}. \tag{12.2}$$

JC has been a well established measure in the information retrieval and data mining community and quantifies the probability that a common neighbour of a pair of nodes would be selected if the selection is performed randomly from the union of sets of neighbors $\Gamma(u)$ and $\Gamma(v)$ [193].

The weighted preferential attachment (PA) is according to [157] :

$$PA(u,v) = \sum_{a\in\Gamma(u)} w_{au} * \sum_{b\in\Gamma(v)} w_{bv}. \tag{12.3}$$

PA considers only the degrees of two nodes, while weighted PA also considers their weights. It has been shown that PA governs the evolving of scale-free networks [204, 205].

The weighted Adamic-Adar (AA) adapted from [157], according to the original unweighted definition in [199], is :

$$AA(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} \frac{w_{uz} + w_{vz}}{\log\left(1 + \sum_{a\in\Gamma(z)} w_{za}\right)}. \tag{12.4}$$

Measure	Notation	Equation
Weighted common neighbors	CN	$CN(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} (w_{uz} + w_{vz})$
Weighted Jaccard coefficient	JC	$JC(u,v) = \frac{\sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} (w_{uz}+w_{vz})}{\sum_{a\in\Gamma(u)} w_{au} + \sum_{b\in\Gamma(v)} w_{bv}}$
Weighted preferential attachment	PA	$PA(u,v) = \sum_{a\in\Gamma(u)} w_{au} * \sum_{b\in\Gamma(v)} w_{bv}$
Weighted Adamic-Adar	AA	$AA(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} \frac{w_{uz}+w_{vz}}{\log\left(1+\sum_{a\in\Gamma(z)} w_{za}\right)}$
Weighted resource allocation index	RA	$RA(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} \frac{w_{uz}+w_{vz}}{s_z}$
Selectivity	SE	$SE(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} \frac{s_z}{deg(z)}$
Inverse selectivity	IS	$IS(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} \frac{deg(z)}{s_z}$

Table 12.1: **Summary of link prediction measures.** In particular, u,v,z,a,b are nodes, w are weights on the links, s_u is the strength, $deg(u)$ is the degree, and $\Gamma(u)$ is the set of neighbors of the node u.

AA ranks the common neighbors with a smaller degree more heavily, and punishes the common neighbors with a higher degree.

The weighted resource allocation index (RA) where s_z is the strength of node z is defined in [156] as :

$$RA(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} \frac{w_{uz}+w_{vz}}{s_z}. \tag{12.5}$$

RA punishes the common neighbors with higher strength more heavily and promotes the ones with lower strength. It assumes the amount of resources that the node can share in its neighbourhood. RA was initially defined as $\sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} \frac{1}{s_z}$ [206]. Since Lü and Zhou [156] report that the unweighted resource allocation index sometimes performs better then the weighted, we decided to use the unweighted variant of RA. The unweighted RA is governed by the same underpinning idea as selectivity and this will allow better insights into a comparative analysis of RA with two newly proposed measures.

Selectivity (SE) is defined as

$$SE(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} \frac{s_z}{deg(z)} \tag{12.6}$$

where $deg(z)$ is the degree and s_z is the strength of node z. Selectivity, originally proposed by Masucci and Rogers [11], promotes the nodes with high strength and low degree, and depresses the high degree nodes. The same governing principle is exploited in the Adamic-Adar and resource allocation index. Since resource allocation has been very successful in link prediction we were motivated to test inverse selectivity as the potential link prediction measure as well.

Inverse selectivity (IS) is defined as a degree of node z divided by it's strength :

$$IS(u,v) = \sum_{z\in\Gamma(u)\cap z\in\Gamma(v)} \frac{deg(z)}{s_z}. \tag{12.7}$$

Resource allocation index, selectivity and inverse selectivity are all computationally undemanding. In order to summarize the seven link prediction measures we systematically list their notation and the corresponding equations in Table 12.1.

12.3.1 Evaluation Metrics

In order to test the performance of weighted similarity measures we need to establish a testing set of links E_P which is used as a golden standard for evaluation. When we usually use a hold-out strategy

for the construction of the test set it holds that the intersection of the training E_T and testing E_P sets is empty $E_T \cap E_P = \emptyset$ and that $E_T \cup E_P = E$. However, in our case we followed different principles for the construction of the testing set. The data is divided into four longitudinally growing subsets, meaning that each of the three training sets is a subset of the testing set.

The link prediction can be evaluated by many different scores as elaborated in [158]. In this work we use: precision, F1 score and the area under the receiver operating characteristic curve (AUC).

The link prediction precision P is the ratio between the number of correctly predicted links and the total number of predicted links - the number of true positives ($|TP|$) divided by the number of true positives and false positives ($|TP| + |FP|$) [158] as:

$$P = \frac{|TP|}{|TP| + |FP|}. \tag{12.8}$$

The F1 score is a standard measure for evaluation in information retrieval tasks and is calculated as the harmonic mean of precision P and recall R:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} = \frac{2 \cdot |TP|}{2 \cdot |TP| + |FP| + |FN|} \tag{12.9}$$

where recall is calculated as a fraction of true positives ($|TP|$) over the number of true positives and false negatives ($|TP| + |FN|$).

The area under the receiver operating characteristic curve (AUC) represents the performance trade-off between the true positive rate against the false positive rate [158, 207]. The receiver operator characteristic curve connects the points corresponding to the pairs of true positive and false positive rates obtained for different decision boundaries. The true positive rate is defined as the fraction of actual positive cases over all positive cases as *correct positives/total positives* or $|TP|/(|TP| + |FN|)$. The false positive rate is the fraction of actual negative cases that are misclassified as positives over all negative cases as *incorrect negatives/total negatives* or $|FP|/(|TN| + |FP|)$. The AUC is calculated as the area under the receiver operating characteristic curve and has values between 0 and 1. The AUC value of 0.5 is a random prediction and higher values are achieved for better models. Hence, the value of 1 represents the score of the perfect model (classifier).

The comparison of different measures for link prediction on several datasets using three evaluation metrics simultaneously amounts to the problem of comparing multiple classifiers over multiple datasets. In order to provide a better insight into the obtained results, we introduce the rank diagrams proposed by Demšar [208]. The rank diagrams position the best value on the left (1st rank) and the worst on the right side, while others are ranked in between. The groups of scores which are not significantly different are connected with the line below the x-axis. The scores (average ranks) are significantly different, if their difference is above the threshold value obtained using the Nemenyi post-hoc test: the threshold is referred to as critical distance CD, calculated as $CD = q_\alpha \sqrt{\frac{K(K+1)}{6N}}$ where q_α is based on Studentized range statistic, K is the number of models (classifiers), and N is the number of measurements (datasets). The critical distance value is depicted on the ranking diagram using a line above the x-axis (labeled CD). All rank diagrams are generated for the Nemenyi test with p-values below 0.05. Figure 12.1 shows an example of the rank diagram. The source code and the explanation of the rank diagrams is available at the Orange Data Mining webpage of the Bioinformatics Lab at the University of Ljubljana.

12.3.2 Datasets

For the link prediction task we exploited two Twitter datasets: the first consists of extracted tweets using the Twitter API (referred to as emo-net) and the second consists of the Sentiment140 corpus with carefully annotated tweets according to their polarity [148] (referred to as SC).

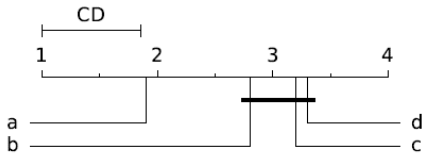


Figure 12.1: **This ranking diagram shows the average ranks for** 4 **models (methods, classifiers): a, b, c and d.** The best ranked (the best performing) model **a** is at the leftmost position, while the worst performing model **d** is ranked at the rightmost position. Others are in the middle according to the achieved rank (measured performance value). The line below shows that the difference between models **b**, **c** and **d** is not statistically significant.

In the emo-net corpus, we extracted four sets of tweets in the English language according to the following search criteria: a) tweets associated to immigrant and war related events (e.g. terrorist, terrorism, ISIS, etc.); b) tweets containing negatively polarized words (e.g. anger, fear, hate, etc.); c) tweets associated to pets (e.g. puppy, kitty, etc.) and d) tweets containing positively polarized words (e.g. joy, happiness, happy, etc.). We will refer to the networks constructed from these sets of tweets respectively as: a) emo-net^a, b) emo-net^b, c) emo-net^c and d) emo-net^d. The four search criteria are selected in order to ensure consistency with the positively or negatively annotated polarity of tweets in the SC dataset, and to keep the data used for the experimental set-up comparable.

The second corpus, SC, consists of four datasets extracted from the SC's training data as follows: a) the first 10,000 negatively polarized tweets, b) the first 10,000 positively polarized tweets, c) the first 100,000 negatively polarized tweets and d) the first 100,000 positively polarized tweets. We will refer to these datasets respectively as: a) $SC^{10^4}_{neg}$, b) $SC^{10^4}_{pos}$, c) $SC^{10^5}_{neg}$ and d) $SC^{10^5}_{pos}$. The *SC* dataset prepared in 2009 is available at http://help.sentiment140.com/for-students/.

Both corpora were subject to the same data-cleaning procedure of stopwords' removal and tokenization at the white spaces in tweets. Table 12.2 summarizes the content of the eight datasets of the English tweets. It is worth noticing that the first six datasets are approximately of the same size (counted in the number of tweets). Also, SC^{10^4} datasets are proper subsets of SC^{10^5} datasets respectively.

For the data preparation we use Python in combination with the Python Twitter Tools package, which provides an easy-to-use interface for the official Twitter API. The extraction during February 2016 resulted in approximately 10,000 tweets for each of the four different datasets, constructing a corpus of 39,882 tweets in total.

The raw emo-net dataset is available at http://langnet.uniri.hr/resources.html.

12.3.3 Network Construction

The language networks construction principle arises from the very nature of the text [11, 209, 210]. The co-occurrence relation in language networks is established between linguistic units within a sentence (here tweet), where the direction of a link reflects the words' sequencing and weight on the link reflects the frequency of word-pairs mutual appearance - weight is the number of tweets in which two words co-occur. For the link prediction task we construct all the networks as undirected and weighted.

First we construct the networks from all the words in the tweets. From emo-net datasets we extract the top 200 most frequent words and extend the list with explicit keywords used for the extraction of tweets (e.g. joy, puppy, anger,...). A link between two nodes is established if these two words co-occur in the same tweet. For the SC datasets we retain the same principles of extracting

Dataset	Number of				
	tweets	words	diff.words	hashtags	diff.hashtags
emo-net[a]	9987	169045	26528	7967	1592
emo-net[b]	9958	151216	25013	1859	985
emo-net[c]	9946	137291	26953	2576	1522
emo-net[d]	9991	143516	31983	3987	2092
$SC^{10^4}_{neg}$	10000	135751	27056	185	151
$SC^{10^4}_{pos}$	10000	130531	30441	183	158
$SC^{10^5}_{neg}$	100000	1349841	150611	1843	1087
$SC^{10^5}_{pos}$	100000	1283953	175722	2394	1324

Table 12.2: **Eight datasets of English tweets considered in this work.** In the emo-net dataset the tweets are extracted according to positive and negative search criteria (e.g. fear, hate, joy, puppy, etc.), while in the SC dataset tweets are selected from already annotated positive and negative polarity of the tweets [148]. The number of different words and the number of different hashtags exclude repetitions, while the number of words and hashtags are the total values including repetitions. We note that the SC^{10^4} datasets are proper subsets of the larger SC^{10^5} datasets.

the top 200 most frequent words and network construction. Next we construct hashtag networks. From both datasets we extract the top 200 most frequent hashtags, and a link is established between hashtags co-occurring in a tweet. Note that the number of different hashtags in $SC^{10^4}_{neg}$ and $SC^{10^4}_{pos}$ is below 200 (see values listed in Table 12.2), so we use the available top-frequent set. The principle of using the top 200 most frequent words (hastags) provides the best trade-off between computation time and link prediction results. Still, in order to test whether using the larger top set contributes to the change in the results we also probe the top 500 extracted hashtags in the $SC^{10^5}_{pos}$ dataset.

Finally, for each of the eight datasets for all-words and for hashtags respectively, we create subnetworks by adding 25%, 50% and 75% of the links, while the entire network of 100% links serves as the baseline for evaluation. The subnetworks preserve the temporal aspect of network construction process, since links are added according to the time of creation captured in the tweet's timestamps. In other words, we construct networks from the sorted list of tweets (from the oldest to the newest) .

To summarize, in total we construct 64 networks (32 based on all-words and 32 based on the hashtags in the tweets), systematically using 25%, 50%, 75% and 100% of the links. 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 [8].

12.3.4 Link Prediction

The link prediction process is the same across all networks (25%, 50% and 75% of the links), regardless of whether the networks are constructed for the co-occurrence of all-words or hashtags in tweets . First, for each dataset we establish the test dataset E_P as a full network with 100% of the links. Then the link candidates are selected from all non-existing links in the current network (25%, 50% and 75%) and ranked according to the assigned value of the link prediction measures. Then we cut off the top n potential links, where n is the total number of new links in the respective testing network, and construct a candidate set. The full set of valid (true positive) future links is generated from the 100% network. Then, two sets (predicted and real links - true positive) are used for the evaluation in terms of precision, the F1 score and the area under the receiver operating characteristic curve (AUC).

12.4 Results

In this Section, we show all the results needed to communicate the main message of our research, while additional results are provided in the S1 Text, together with the definition of a standard set of network measures used for exploring the structure of networks.

12.4.1 Link Prediction Results in All-word Networks

The link prediction results in networks constructed from all the words in tweets are presented in Figure 12.2 for the emo-net dataset, while Figure 12.3 shows the results for the SC dataset. In both figures the results are contrasted between precision, the F1 score and the area under the receiver operating characteristic curve (AUC). It can be observed that the F1 score and precision follow the same regularities i.e. exhibit decreasing values from the 25% to 75% networks regardless of the dataset. In emo-nets the weighted preferential attachment (PA) is systematically under-performing while the weighted Jaccard coefficient (JC) slightly deteriorates in the SC^{10^4} datasets. The achieved results are in a favor of larger datasets. Also the difference between the F1 score and precision is lower in the SC datasets, especially in SC^{10^5} and link prediction performance increases with the size of the data used. AUC exposes no substantial variability over different datasets, improvement is only noticed in larger in datasets (SC^{10^5}) regardless of the link prediction measure. From the presented results it is difficult to judge about the performance of the tested link prediction measures, therefore the analysis of ranking of seven link prediction measures follows.

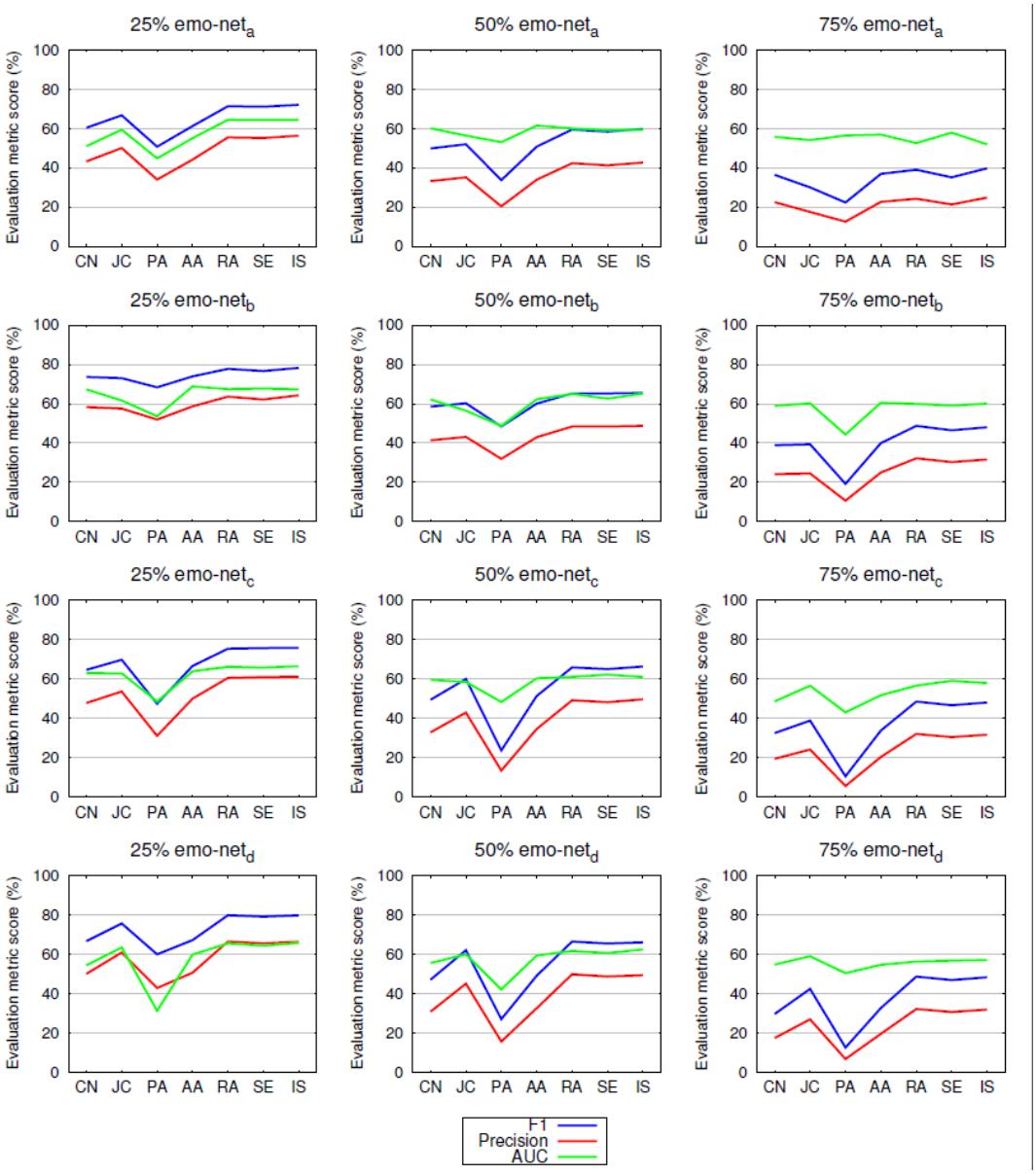


Figure 12.2: **Link prediction in 25%, 50% and 75% of the links in networks constructed from all the words in tweets of the emo-net^a, emo-net^b, emo-net^c and emo-net^d datasets.** Shown are the evaluation metric scores (see legend), namely the F1 score, the precision, and the area under the receiver operating characteristic curve (AUC), as obtained for seven different link prediction measures, namely common neighbors (CN), the Jaccard coefficient (JC), preferential attachment (PA), Adamic-Adar (AA), the resource allocation index (RA), selectivity (SE) and inverse selectivity (IS). The values of the F1 score and of precision are decreasing with the longitudinal growth of the networks (from 25% to 75%), while the AUC does better at retaining values regardless of the used percentage of links. The PA link prediction measure exposes the lowest link prediction potential on the emo-net dataset, this is regardless of the evaluation metrics used. See Table 12.2 and the main text for details.

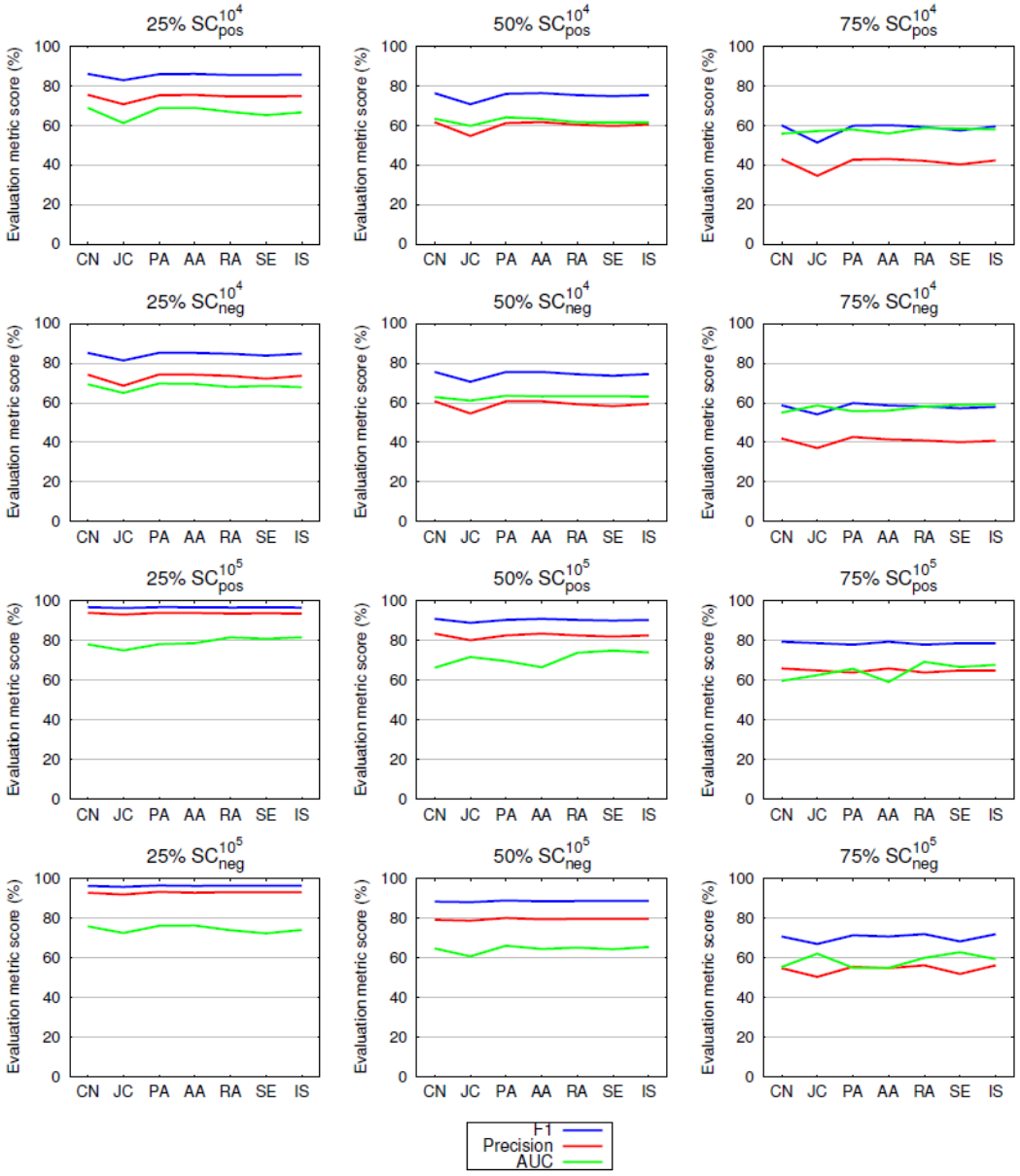


Figure 12.3: **Link prediction in 25%, 50% and 75% of the links in networks constructed from all the words in tweets of the $SC^{10^4}_{neg}$, $SC^{10^4}_{pos}$, $SC^{10^5}_{neg}$ and $SC^{10^5}_{pos}$ datasets.** Shown are the same quantities as in Figure 12.2. Here too the values of the F1 score and of precision are decreasing with the longitudinal growth of the networks (from 25% to 75%), while the AUC does better at retaining values regardless of the percentage of links used. It can also be observed that larger networks yield better link prediction measures. See Table 12.2 and the main text for details.

In Figure 12.4 we show rank diagrams for the F1 score (left) and the area under the receiver operating characteristic curve (AUC) (right) for the 25% (top), 50% (middle) and 75% (bottom of the figure) networks from all-words in tweets over all datasets.

Rankings between precision (see data in S1 Text) and the F1 score are preserved for the 25% and 75% networks, while the rankings with AUC exhibit a different trend. Inverse selectivity (IS) is at the highest rank according to the F1 score, while AUC ranks the resource allocation index at the top position. Additionally, we consider the average overall rank across all networks (25%, 50% and 75%) of link prediction measures which positions at the top three places IS, AA, RA (according to the F1 score evaluation) and RA, SE and IS (according to the AUC evaluation).

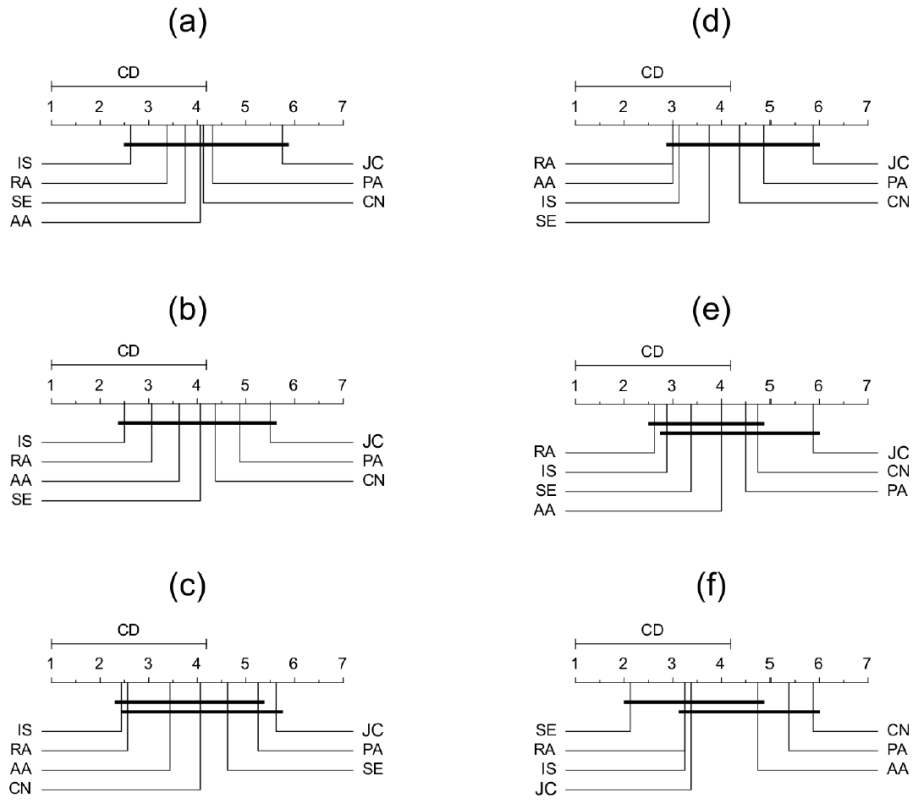


Figure 12.4: **Ranking diagrams based on networks constructed from all the words in tweets for the seven link prediction measures used in this research.** Namely for common neighbors (CN), the Jaccard coefficient (JC), preferential attachment (PA), Adamic-Adar (AA), the resource allocation index (RA), selectivity (SE) and inverse selectivity (IS). Rankings according to the F1 score are presented on the left for 25% (a), 50% (b) and 75% (c), while rankings according to the area under the receiver operating characteristic curve (AUC) are presented on the right for 25% (d), 50% (e) and 75% (f). The best rank is at the leftmost position and the line below denotes measures which are not significantly different (Nemenyi test with p-values of 0.05).

12.4.2 Link Prediction Results in Hashtag Networks

Next we analyze the difference between the hashtags' networks compared to the all-words networks. Regardless of the tested measures or corpora, the results are only changed slightly–mainly deteriorated but in some cases also slightly improved .

Figures 12.5 and 12.6 compare the area under the receiver operating characteristic curve (AUC) values of the all-words and hashtags networks. If we consider the F1 score as an evaluation metric on smaller emo-net datasets, the results of all-words over the respective hashtag networks are improved by 13-37% (for the 25% networks); 11-30% (50% networks) and 8-21% (75% networks). On the SC dataset the results of the all-words' networks are better by: 38-50% (25%); 43-53% (50%) and 35-54% (75%). In terms of AUC the observed differences are in general smaller: for emo-net up to 30% (25% networks); 19% (50%) and 22% (75%) and for the SC datasets up to 20% (25%); 15% (50%) and 25% (75%).

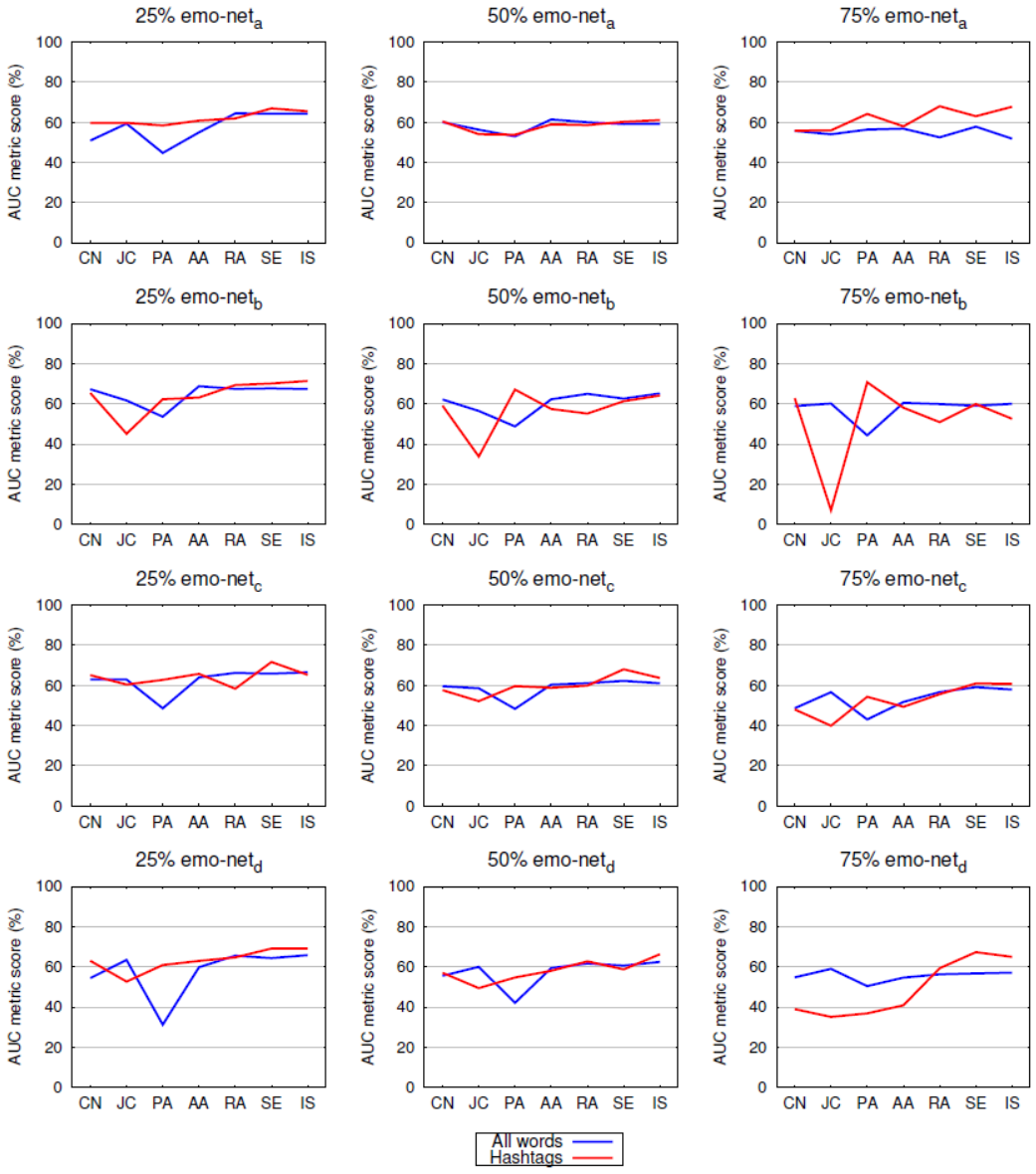


Figure 12.5: **Link prediction in 25%, 50% and 75% of links in networks constructed from all the words and from hashtags (see legend) in tweets of the emo-net^a, emo-net^b, emo-net^c and emo-net^d datasets.** Shown is the area under the receiver operating characteristic curve (AUC), as obtained for seven different link prediction measures, namely common neighbors (CN), the Jaccard coefficient (JC), preferential attachment (PA), Adamic-Adar (AA), the resource allocation index (RA), selectivity (SE) and inverse selectivity (IS). See Table 12.2 and the main text for details.

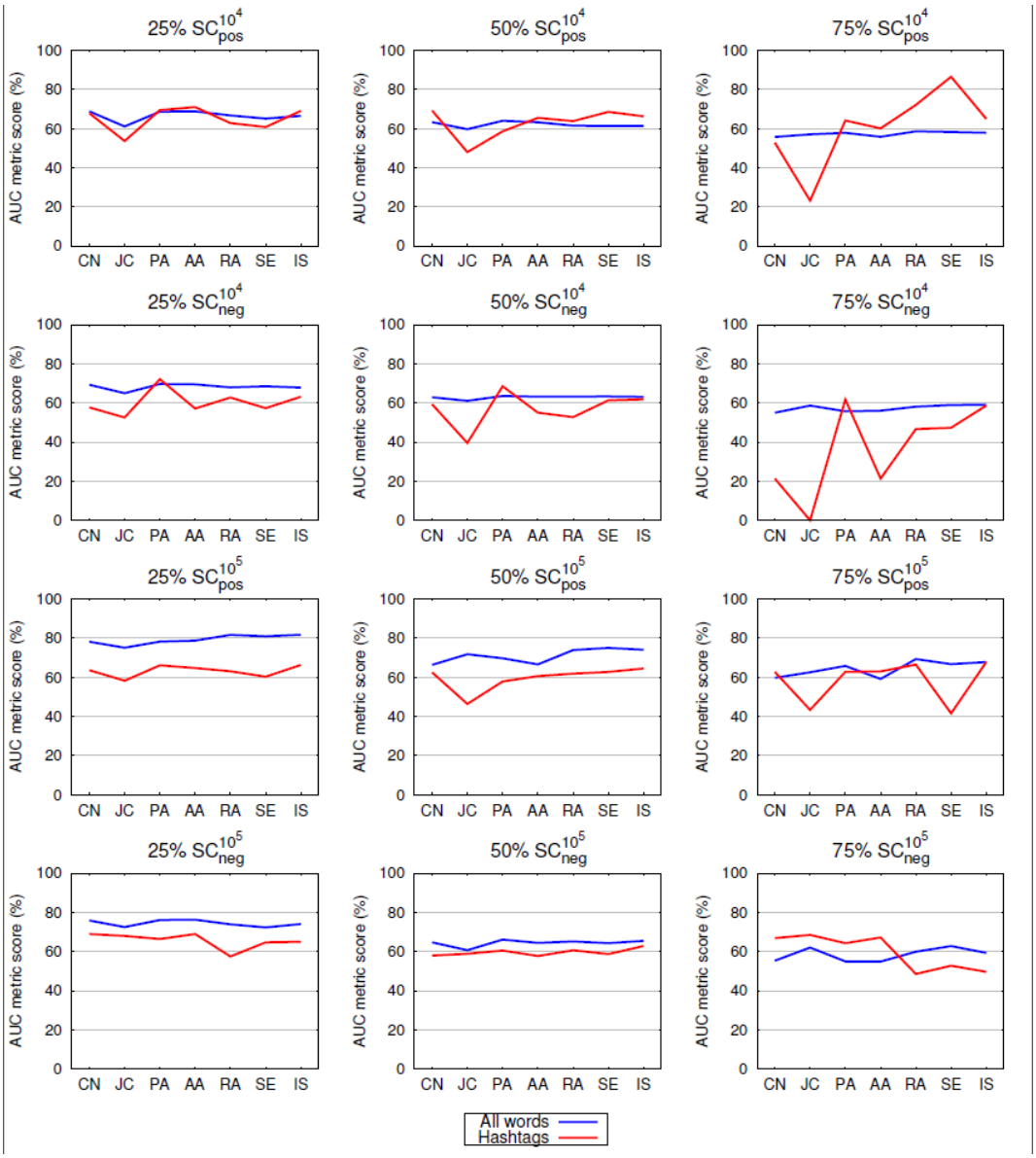


Figure 12.6: **Link prediction in 25%, 50% and 75% of the links in networks constructed from all the words and from hashtags (see legend) in tweets of the $SC^{10^4}_{neg}$, $SC^{10^4}_{pos}$, $SC^{10^5}_{neg}$ and $SC^{10^5}_{pos}$ datasets.** Shown are the same quantities as in Figure 12.5. See Table 12.2 and the main text for details.

Finally, the ranks are presented in Figure 12.7 for the hashtags' networks of the 25%, 50% and 75% of the links for the F1 score (left) and AUC (right) respectively. The rank analysis reveals that the F1 score and AUC are interchanging Adamic-Adar, selectivity and inverse selectivity at the highest positions. The top overall average ranks achieved for the F1 score and AUC on the hashtags are: IS, AA, PA and IS, SE, PA respectively.

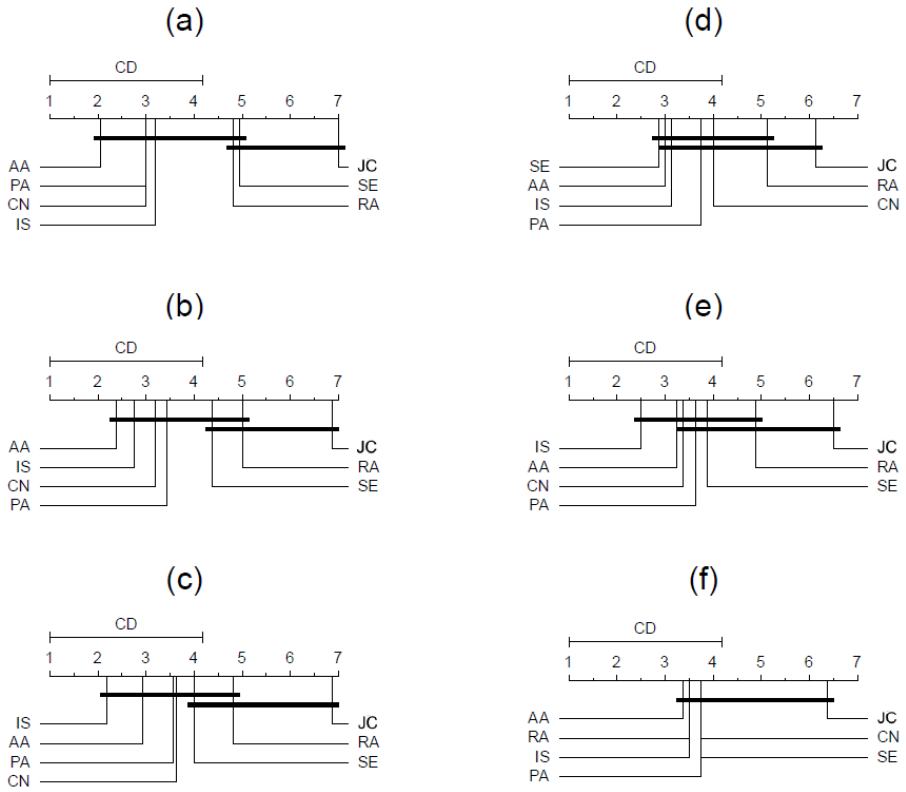


Figure 12.7: **Ranking diagrams based on networks constructed from the hashtags in tweets for the seven link prediction measures used in this research.** Namely for common neighbors (CN), the Jaccard coefficient (JC), preferential attachment (PA), Adamic-Adar (AA), the resource allocation index (RA), selectivity (SE) and inverse selectivity (IS). Rankings according to the F1 score are presented on the left for 25% (a), 50% (b) and 75% (c), while rankings according to the area under the receiver operating characteristic curve (AUC) are presented on the right for 25% (d), 50% (e) and 75% (f). The best rank is at the leftmost position and the line below denotes measures which are not significantly different (Nemenyi test with p-values of 0.05).

Alternative rankings according to different evaluation scores indicate the need for considering different evaluation metrics simultaneously, while using only one metric provides myopic insights into the results. This is strong evidence that multiple evaluation metrics should be considered for the evaluation of link prediction of the future content of tweets. The reported results also suggest that F1 score is a better candidate than precision, so for future research in link prediction in language networks we suggest considering the F1 score and AUC in parallel.

Finally, we test whether the network construction principles of cutting off the top 200 most frequent words (hashtags) influences the obtained results. The construction of the top 500 hashtags' networks follows the same principles except that the cut-off threshold is set to 500 instead of 200. The $SC^{10^5}_{pos}$ dataset was selected due to the sufficient number of different hashtags and the size of 10^5. The results in Figure 12.8 depict the differences between the obtained top 200 and top 500 results in terms of the F1 and AUC scores for the 25%, 50% and 75% hashtags' networks respectively. There

are insignificant differences in the obtained results between the top 200 and the top 500 networks, except for the AUC from the 75% networks. AUC notably deteriorates in SC^{10^4} 500 networks, due to the number of different hashtags below 160.

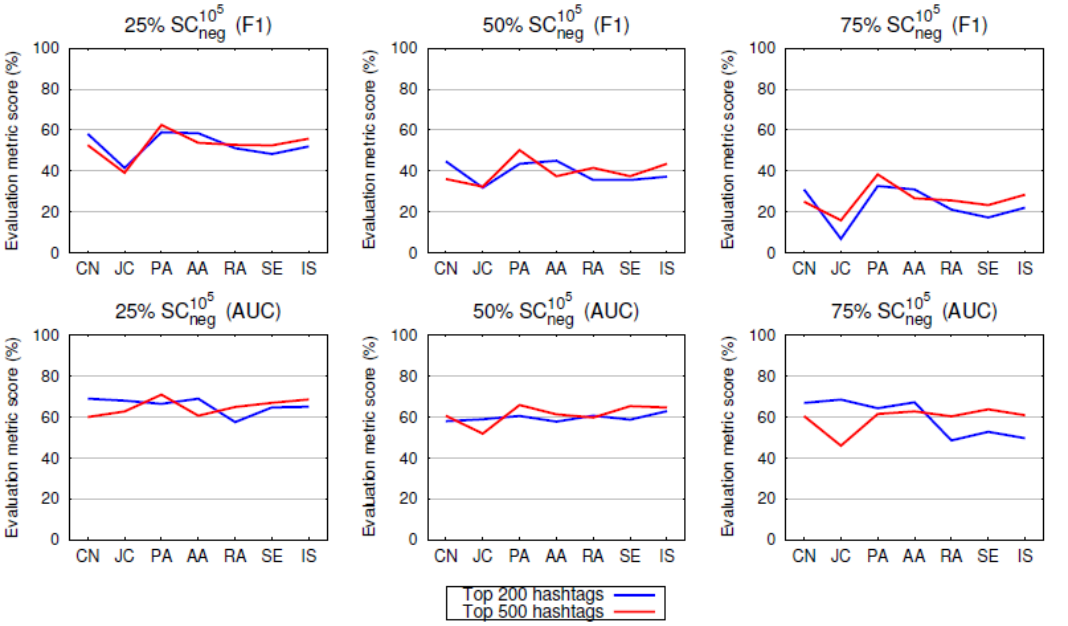


Figure 12.8: **Link prediction in 25%, 50% and 75% of the links in networks constructed from the top** 200 **and top** 500 **hashtags (see legend) in tweets of the $SC^{10^5}_{neg}$ dataset.** The upper row shows the F1 score, while the bottom row shows the area under the receiver operating characteristic curve (AUC), as obtained for the seven different link prediction measures considered in this research.

12.5 Discussion

The trend of decreasing precisions and F1 score values along the 25% to 75% links in networks is present for all-words' and hashtags' networks. In networks created from 25% of the data, many probable links are left out. At the same time the most probable links are the most likely to be predicted and the link prediction measures are the most successful in predicting highly-probable links. With more data in the 50% and 75% networks the majority of highly-probable links are already included in the network, therefore the prediction measure is expected to predict less-probable links, which causes the drop in the prediction precision and the F1 score. At the same time AUC is prone to this effect. Zhao et al. in [197] observe similar problems in the dataset for testing, which they overcome by computing the odds ratio for correcting the prediction results. Following the same principle we plan to introduce the odds ratio into the evaluation of link prediction in language networks.

Regarding the size of the used datasets (10^5 vs 10^4 in SC) we can conclude that more data raise the improvement in the obtained results (as expected) - F1 scores are improved but the values of the area under the receiver operating characteristic curve (AUC) are of the same range and not notably higher. Hence, we can consider the results for the 10^4 size as representative, especially when we regard the network construction principles being the same and resulting from networks of approximately the same size of nodes.

The F1 score and precision values shown in Figures 12.2 and 12.3 exhibit regularities across tested link prediction measures and datasets. The F1 score, calculated as the harmonic mean of

precision and recall, is a more suitable evaluation metric than precision. Hence, we confirm the findings for social follower networks in [191], and for reciprocal follower networks on Twitter in [194] also for language networks constructed from the content of tweets – words and hashtags.

The two newly proposed measures for link prediction selectivity (SE) and inverse selectivity (IS) proved correct, especially IS which is ranked the best in 8 out of 18 cases, AA is the best 5 times, while SE and RA are at the top ranked position twice. In contrary JC occurred 17 times at the lowest rank. This is in accordance with other reported results where the measures which punish the nodes with a higher degree (AA, RA, SE and IS) are overperforming common neighbors, the Jaccard coefficient and preferential attachment in biological, social or technical networks [190, 196, 197]. Due to the achieved scores and low computational cost, we can conclude that selectivity and inverse selectivity should be considered for weighted link prediction, especially when dealing with texts in language networks.

Due to the same construction principles we analyse networks of a similar size, which is reflected on the very comparable results in hashtags to all-words' networks. The network density is high and as expected systematically increasing from the 25% to 100% all-words' networks, while hashtags' networks exhibit some variations, especially in the SC dataset. Murata and Moriyasi in [196] discuss the positive influence of the network density on the performance of the weighted similarity measures, which is also reflected in our results. Next, all the studied networks are characterized by a relatively high average clustering coefficient, a very high average degree and average strength underpinning the efficiency of weighted similarity measures in both words' and hashtags' networks.

The area under the receiver operating characteristic curve (AUC) value of 0.5 is a random prediction – there is no relationship between the predicted values and the truth. An AUC below 0.5 indicates there is a relationship between the predicted values and the truth, but the model is backwards, i.e., predicts smaller values for positive cases. Another way to think of AUC is to imagine sorting the data by predicted values. Suppose this sort is not perfect, i.e., some positive cases sort below some negative cases, then AUC effectively measures how many times you would have to swap cases with their neighbors to repair the sort. Thus, sometimes we obtain a value below 0.5 for the weighted preferential attachment measure. All the networks have an assortativity between -0.02 and -0.52 which characterize the networks from the content of tweets as non-assortative. This is related to preferential attachment indicating that this is not the underlying mechanism for the growth of language networks. Finally, this is reflected in the score of preferential attachment with some AUC values below 0.5.

Link prediction is known to be an unbalanced classification problem and the receiver operating characteristic curves are insensitive to changes in class distributions and therefore insensitive to skewed class distributions [207]. Hence, it is no surprise that AUC metric provides more consistent insights into a measured performance over different datasets. Still, it would be wrong to neglect the F1 score for the evaluation since it provides a different perspective of the results. This is especially important, since we are dealing with text and hashtags. The content of microblogs represented in the form of words and hashtags is important for information representation and information propagation which are of interest in the information retrieval discipline as well. Information retrieval is traditionally oriented towards the F1 score based evaluations. Hence based on our findings we advocate the use of the F1 score and AUC simultaneously. To conclude, we find the introduced rank diagrams as a very useful tool which helps in merging the results of two or more evaluation metrics, and undoubtedly helps in gaining a holistic overview of the link prediction measures' performance over different datasets.

In general hashtag networks exhibit similar characteristics as all-word networks: there is less difference of the AUC values than in terms of the F1 scores; hashtags constantly have lower F1 scores than all-words' counterparts, while AUCs are of the same range. F1 scores are decreasing from the 25% to 75% networks, while AUC expose constant values; and there are no significant

deviations in results on larger datasets. The only salient behaviour is noticed between the number of hashtags in the emo-net and SC datasets: it seems that the more recent tweeting trends rise more systematic (frequent) use of hashtags, which is reflected onto the structural properties of the studied networks. The influence of the distribution of hashtags per tweet is elaborated in [186] where they report about 50% of tweets tagged with one hashtag (dataset collected in 2013), while authors in [153] report around 15% of tweets with one hashtag (dataset collected before 2011). Next, the expansion of the network structure to the top 500 hashtags (Fig 12.8) exhibited no significant improvements. The importance of hashtags is reflected in capturing the semantic context of tweets, and as such are important for the summarization and categorization of the tweets's content. This study is an initially step toward revealing the deeper structural properties of hashtags and will be addressed in our future studies.

12.6 Conclusions

In this work we analysed link prediction based on the local similarity measures on networks constructed from the content of tweets: all-words and hashtags. The main goal of this analysis is to find which measure performs better in the task of predicting the future linking of words and hashtags in the content of tweets, which can be utilized for the propagation of information and opinion in social networks.

Besides five already analysed measures for link prediction in weighted complex networks of common neighbors (CN), the Jaccard coefficient (JC), preferential attachment (PA), Adamic-Adar (AA) and the resource allocation index (RA), we proposed two new measures: selectivity (SE) and inverse selectivity (IS). The experimental results obtained from two corpora of English tweets through the construction of systematically growing subnetworks form the 25%, 50% and 75% of the links and evaluated on the full content of 100% of the links in the network revealed many new findings.

First, the introduced ranking diagrams proved beneficial, as a powerful and straightforward tool for comparing the achieved scores of multiple tested link prediction measures on multiple datasets. The alternative rankings achieved by different evaluation scores (the F1 score and the area under the receiver operating characteristic curve) indicate the need to consider multiple evaluation metrics simultaneously, in order to obtain an unimpeded perspective on the link prediction on Twitter. Second, the two newly proposed measures selectivity (SE) and inverse selectivity (IS) proved efficient, especially IS, which is ranked best in 8 out of 18 cases, AA is the best 5 times, while SE and RA are at the top ranked position twice. In contrast, JC occurred 17 times at the lowest rank. Inverse selectivity is the first choice of measures for the task of predicting the future content of tweets. Third, the hashtags results exhibit similar characteristics as all-words networks, and as such are suitable candidates for the further examination of the content on Twitter within a complex network framework. Besides that, hashtags are able to capture the semantic context of tweets, and as such are important for the summarization and categorization of tweets.

The presented research reveals many possible direction for future studies. The focus of our future research plans is a deeper investigation of hashtag networks, incorporating the prediction of weights on the links and introducing the odds ratio to evaluate weighted link prediction in language networks.

13. Extracting Domain Knowledge by Complex Networks Analysis of Wikipedia Entries

13.1 Abstract

In this Chapter we describe a complex networks analysis of Wikipedia. We construct 10 different networks from Wikipedia entries (articles) related to the chosen domain. The goal of the experiment is to extract domain knowledge in terms of identifying entries that are centrally positioned and entries that are strongly related. We apply complex networks analysis on all acquired networks and examine the networks' structure. We employ centrality measures in order to find centrally positioned entries in the network. Furthermore we identify communities and find which entries are densely connected according to the network structure.

13.2 Introduction

Complex networks exhibit specific topological features, such as high clustering coefficients, small diameters, a power-law degree distribution, community structure, one or several giant components, hierarchical structures, etc. Two important classes of complex networks that can be further differentiated are small-world networks [43] with small distances and high clustering coefficients as main properties and scale-free networks [43] which can be characterized by a power-law degree distribution.

Wikipedia can be modelled as a complex network in a way that Wikipedia entries are nodes, and links between two nodes are established if there is a hyperlink between these two entries. Early attempts to quantify Wikipedia using complex networks analysis were focused only on network structure of linked Wikipedia entries. In [216] Zlatić et al. present an analysis of Wikipedias in several languages as complex networks. They show that many network characteristics (degree distributions, growth, topology, reciprocity, clustering, assortativity, path lengths and triad significance profiles) are common to Wikipedias in different languages and show the existence of a unique growth process. The same authors studied Wikipedia growth based on information exchange in [217]. In [212] an analysis of the statistical properties and growth of Wikipedia is presented. Pemble and Bingol [27] have constructed two complex networks out of English and

German Wikipedia corpora and analyzed conceptual networks in different languages.

The other research direction is focused on content found on Wikipedia and analyses Wikipedia as a (domain) knowledge network. In Fang [213] they first extract a specific domain knowledge network from Wikipedia (specifically, four domain networks on mathematics, physics, biology, and chemistry) and then carry out statistical analysis on these four knowledge networks. Also, they show that MathWorld and Wikipedia Math share a similar internal structure. In [214] Masucci et al. extract the topology of the semantic space and measure the semantic flow between different Wikipedia entries. They further analyze a directed complex network of semantic flow. In [3] the results of semantic language networks analysis are presented in general. Motivated by the second approach that studies Wikipedia as a knowledge network , we wanted to study how the network structure is related to domain knowledge. The goal of our experiment was to extract centrally positioned entries in the network and analyze how these entries are related to domain knowledge and are some more important than other. In the second part of the experiment the task was to extract entries that belong to the same community and check whether they are semantically related.

In our previous research, we have already analyzed Wikipedia as a complex network [31], but by constructing a network of syllables. Also, we examined the structure of Croatian language networks in [21, 50, 72]. In [59, 72] we applied network measures for a keyword extraction task. In all our previous experiments we were focused solely on language structure and this is our first attempt to analyze semantic relations in a network.

In the Section 13.3 we present key measures of complex networks involved in network structure analysis. In the Section 13.4 we describe data sources and network construction principles. In the Section 13.5 we present the results. Finally, the Section 13.6 contains a conclusion and possible directions for future research.

13.3 Network Structure Analysis

In this Section we review some of the most important network measures [67]. Every network has an N number of nodes and a K number of links. The degree of a node i is the number of links with which the node is connected, k_i. Considering the fact that we are working with directed networks, we must specify two types of degrees: the in-degree, k_i^{in}, corresponding to the number of incoming links and the out-degree, k_i^{out}, equal to the number of outgoing links for any particular node i. The average degree of the network is:

$$< k >= \frac{2K}{N}. \tag{13.1}$$

For the directed networks we omit multiplication by 2. In the further equations we assume that the network is directed and that the total possible number of links is equal to $N(N-1)$. For every two connected nodes i and j, the number of connections lying on the path between them is represented as d_{ij}, and so d_i is the average distance of a node i from all other nodes, and it is obtained by:

$$d_i = \frac{\sum_j d_{ij}}{N}. \tag{13.2}$$

For the next two measures, if a network contains more than one component, we consider the largest component . The average shortest path length between every two nodes in a network is:

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}. \tag{13.3}$$

And the maximum distance results in the network diameter, D:

$$D = max_i d_i. \tag{13.4}$$

The clustering coefficient is a measure which defines the presence of connections between the nearest neighbours of a node. And so, c_i (the clustering coefficient) of a node is a fraction between the number of edges E_i that exist between node i and the total possible number of edges within the neighbourhood of the node i:

$$c_i = \frac{2E_i}{k_i(k_i-1)}. \tag{13.5}$$

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

$$C = \frac{1}{N}\sum_i c_i. \tag{13.6}$$

Density of a network is a measure of network cohesion defined as the number of observed links divided by the number of total possible links:

$$d = \frac{K}{N(N-1)}. \tag{13.7}$$

Degree centrality of a node i is the degree of that node. It can be normalised by dividing it by the maximum possible degree $N-1$:

$$dc_i = \frac{k_i}{N-1}. \tag{13.8}$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Let σ_{jk} be the number of shortest paths from node j to node k and let $\sigma_{jk}(i)$ be the number of those paths that pass through the node i. The normalised betweenness centrality of a node i is given by:

$$bc_i = \frac{\sum_{i\neq j, i\neq k}\frac{\sigma_{jk}(i)}{\sigma_{jk}}}{(N-1)(N-2)}. \tag{13.9}$$

Closeness centrality is defined as the inverse of farness, i.e. the sum of the shortest distances between a node and all other nodes. Let d_{ij} be the shortest path between nodes i and j. The normalised closeness centrality of a node i is given by:

$$cc_i = \frac{N-1}{\sum_{i\neq j} d_{ij}}. \tag{13.10}$$

Modularity measures the quality of the network partition in the communities. The modularity of a network partition is a scalar value between -1 and 1 that measures the density of links inside communities as compared to links between communities. Let e_{ij} be the fraction of edges in the network that connect vertices in group i to those in group j, and let $a_{ij} = \sum_j e_{ij}$. Then the modularity can be calculated using following equation :

$$Q = \sum_{i=1}^{k}(e_{ii} - a_i^2). \tag{13.11}$$

The degree assortativity coefficient measures the tendency of nodes in a network to connect to nodes similar to themselves. The coefficient lies between -1 and 1 and it is quantified via the Pearson correlation. Positive r values indicate a correlation between similar-degree nodes. Let q_k and q_j be the distribution of the degree of out-edges that do not connect to the other node in

question, e_{jk} the joint probability distribution of q_k and q_j, and σ_q^2 the variance of the distribution. Then we can calculate the assortativity coefficient using the following equation :

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}. \tag{13.12}$$

On the meso-scale level complex networks analysis includes a community detection task [167]. Communities, in this sense, are groupings of densely interconnected nodes within a network. In other words, nodes in a community have a greater amount of connections amongst each other than with other nodes in the network . Several algorithms are used for community detection such as hierarchical clustering, Girvan-Newman's algorithm, minimum-cut method, etc . One of the most efficient is the Louvain method [211], a greedy optimization method that optimizes the modularity of a network's partitions. The number of communities (N_c) represents the amount of such groupings found within a network.

13.4 Network Construction

For the purpose of our experiment we collect entries from Wikipedia and construct networks related to the domain. Our intention was to construct two types of networks: level 2 networks and level 4 networks. We construct level 2 networks by starting with a chosen seed entry (e.g. "Complex network" or "Data"), storing all the hyperlinks to related entries from the seed entry's text (level 1) and proceeding to extract the hyperlinks from all the entry pages taken from the original entry (level 2). Analogously, we construct level 4 networks by taking the first 10 hyperlinks from a given entry page and proceeding to repeat the task three times, arriving at level 4. We limit the hyperlinks to the first 10 due to the computational complexity at the same time having in mind that the most general hyperlinks are usually at the beginning of the entry's text.

Therefore, the first task is the construction of a web scraping program which would extract hyperlinks from a Wikipedia entry's text. The hyperlinks are extracted using a Python package for HTML parsing called Beautiful Soup which parses the HTML structure of a given HTML document into a parse tree. By navigating the tree we locate the tag ID which corresponds to article content ("mw-content-text") and proceed to extract the hyperlinks which themselves are found within paragraph (<p>) tags and finally inside link (<a>) tags in that section of the page. Finally, each network is stored in an edge list in the following format: "entry title" \t "linked entry title". We had some difficulties with processing non-ASCII script and hyperlinks that were not connected to other documents (citations, in-page references, etc.), but we managed to avoid those by checking the data during the extraction process.

In our directed network , each entry's title represents a node and it is connected to other entries hyperlinked in its text, again represented as network nodes. We construct a total of 10 domain networks for five chosen seed entries: "Byte", "Complex network", "Computer science", "Data" and "Programming language". The naming scheme includes the level of a specific network in its name (e.g. the level 2 network for "Byte" is **BT2**). Since we consider unweighted networks, we dismiss double links. This, along with the fact that some entries do not contain 10 hyperlinks resulted in our level 4 networks having less than 10^4 expected edges. We use Python and the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [8].

The various visualizations of the networks are done using Gephi [29], an open-source network analysis and visualization package written in Java. The following visualization (Figure 13.1) represents a level 2 network constructed around the "complex network" Wikipedia entry. We loaded the edge list into Python, ran the Yifan-Hu layout algorithm, correlated the label size with the corresponding node's betweenness centrality measure and coloured clusters according to their respective modularity class.

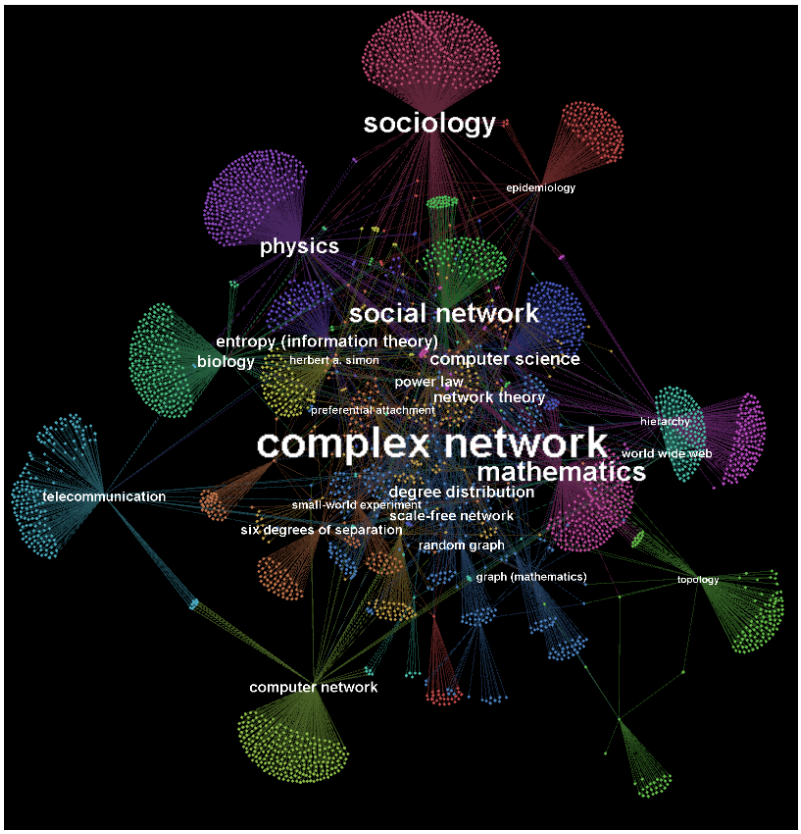


Figure 13.1: CN2 network visualization.

13.5 Results

In this Section we present the results of our measuring described in Section 13.3, such as average degree $< k >$, average path distance L, diameter D, average clustering coefficient C, density d, modularity Q, number of communities (N_c) and degree assortativity coefficient r. We also present the most central nodes (according to the three centrality measures) and communities in networks detected by using the Luvain algorithm.

In Table 13.1 we present estimated global network measures. There are certain differences between measures for level 2 and 4 which are evident upon closer inspection. For instance, level 4 networks have significantly larger average path lengths, diameters, assortativity coefficients, often a significantly larger number of detected communities and slightly larger average degrees. The modularity measure and density are comparable between the two, whilst level 2 networks show larger clustering coefficients.

For comparison with random networks , the table also includes two measures for equivalent random networks (Erdös-Rényi random graphs) - the average shortest path length $(L_{ER} = lnN/ln < k >)$ and the average clustering coefficient $(C_{ER} =< k > /N)$. The results show that the complex networks we have constructed have a significantly higher average clustering coefficient than their Erdös-Renyi random graph counterparts . This, in addition with a relatively small average shortest path length L led us to conclude that we are dealing with small-world networks as described by Watts and Strogatz [216]. For the purposes of this comparison we treat the networks as undirected .

Moreover, a distinctly high modularity coefficient Q (higher than 0.7 in all but one network, as visible in Table 13.1) shows a clear tendency towards community clustering of nodes present in the networks. We did not observe any strict rule governing community size across networks, although level 2 networks have an understandably smaller N_c which we contributed to the very construction

Measure	"Byte"		"Complex network"		"Data"	
	BT2	BT4	CN2	CN4	CS2	CS4
Number of nodes (N)	3945	3632	3405	3070	12881	3630
Number of edges (K)	5112	5611	4132	5008	18852	5851
Average degree ($< k >$)	1.296	1.545	1.214	1.631	1.464	1.612
Avg. shortest path (L)	3.693	6.834	3.198	9.218	3.417	6.277
Avg. shortest path (L_{ER})	8.6938	7.26622	9.168	6.791	8.8088	7.0023
Diameter (D)	9	15	6	22	7	14
Average clustering coefficient (C)	0.06	0.021	0.043	0.024	0.074	0.019
Average clustering coefficient (C_{ER})	0.00066	0.00085	0.00071	0.00106	0.00023	0.00089
Density (d)	0.0003	0.00042	0.00035	0.00053	0.00011	0.00044
Modularity (Q)	0.778	0.776	0.794	0.763	0.725	0.771
Number of communities (N_c)	17	32	17	21	23	27
Degree assortativity coefficient (r)	-0.592	-0.048	-0.521	0.021	-0.491	-0.028

Measure	"Computer science"		"Programming language"	
	DT2	DT4	PL2	PL4
Number of nodes (N)	2297	3658	7467	3965
Number of edges (K)	2630	5531	13933	6215
Average degree ($< k >$)	1.145	1.512	1.145	1.612
Avg. shortest path (L)	3.086	6.369	3.127	6.277
Avg. shortest path	9.341	7.4144	10.764	7.078
Diameter (D)	7	14	6	22
Average clustering coefficient (C)	0.043	0.019	0.082	0.021
Average clustering coefficient (C_{ER})	0.0010	0.00083	0.00031	0.00081
Density (d)	0.00049	0.00041	0.00025	0.0004
Modularity (Q)	0.828	0.779	0.594	0.78
Number of communities (N_c)	18	31	19	30
Degree assortativity coefficient (r)	-0.561	-0.048	-0.468	-0.059

Table 13.1: Global measures for level 2 and level 4 networks of "Byte", "Complex networks", "Data", "Computer science" and "Programming language" Wikipedia's entries.

principle as described in Section 13.4.

After the analysis on the global level, we analyse the networks on the local level in terms of centrality measures. Tables 13.2 and 13.3 show lists of top ten entries according to the three centrality measures for the two seed entries: "Computer science" and "Programming language". We analyse the degree centrality, betweenness centrality and closeness centrality. For the degree centrality we treated the network as undirected. For each centrality measure and domain there are two lists of entries, one for level 2 networks and another for level 4 networks. We noticed that the lists for level 2 networks consist of entries that are semantically related to the seed entries ("Computer science" or "Programming language") in a way that might be ascribed as belonging to a hierarchy. This is especially evident for the closeness centrality measure. For example, the list of top ten entries according to the closeness centrality for the seed entry "Computer science" contains other scientific domains (theoretical computer science, mathematics, artificial intelligence, physics, engineering) and for the seed entry "Programming language", list contains some prominent programming languages (C, Java, Perl, Python, C++).

Degree centrality		Betweenness centrality	
CS2	CS4	CS2	CS4
human	mathematics	**computer science**	computer science
university of Cambridge	cell (biology)	computer	information
philosophy	computer science	**mathematics**	protein
industrial revolution	computer	**artificial intelligence**	science
G. W. Leibniz	information	**philosophy**	algorithm
physics	protein	human	logic
eletrical engineering	organism	Gottfried Wilhelm Leibniz	organism
artificial intelligence	dna	algorithm	cell (biology)
mathematics	computer program	**theoretical comp. science**	computing
Alan Turing	philosophy	**physics**	mathematics

Closeness centrality	
CS2	CS4
computer science	computer science
mathematics	information
theoretical comp. science	science
artificial intellgience	mathematics
philosopy	ancient greek
physics	latin
human	computing
G. W. Leibniz	algorithm
engineering	bit

Table 13.2: Top ten entries in the "Computer science" networks **(CS2, CS4)** regarding the three centrality measures: degree centrality, betweenness centrality and closeness centrality.

In the second part of the experiment we analyse communities in all 10 networks in order to explore which entries are grouped together. Figure 13.2 shows most significant entries from the CS2 network grouped into communities. Different communities are presented in different colours. For example, entries related to the mathematics domain (*mathematics, number, set, function, real number,* etc.) are in the red-coloured community; entries related to the computer science domain (*computing, algorithm, compiler,* etc.) are in the orange-coloured community; entries that are related to the biology domain (*cell, organism, gene*, etc.) are in the light-orange coloured community

Degree centrality		Betweenness centrality	
PL2	PL4	PL2	PL4
history of computing-hardware	mathematics	prog. language	prog. language
internet	computer	computer	computer
english language	computer science	**C** (prog. language)	software engineering
computer	physics	compiler	computing
C (prog. language)	set (mathematics)	english language	computer science
python (prog. language)	greek language	computer program	algorithm
University of Manchester	logic	internet	message
perl	language	**perl**	communication
prog. language	central processing unit	**python** (prog. language)	machine
php	electronics	**java** (prog. language)	function (mathematics)

Closeness centrality	
PL2	PL4
prog. language	prog. language
C (prog. language)	ancient greek
computer prog.	computer
java (prog. language)	mathematics
perl	arithmetic
compiler	science
omputer program	greek language
python (prog. language)	physics
control flow	latin
C++	computer science

Table 13.3: Top ten entries for the "Programming Language" networks **(PL2, PL4)** regarding the three centrality measures: degree centrality, betweenness centrality and closeness centrality.

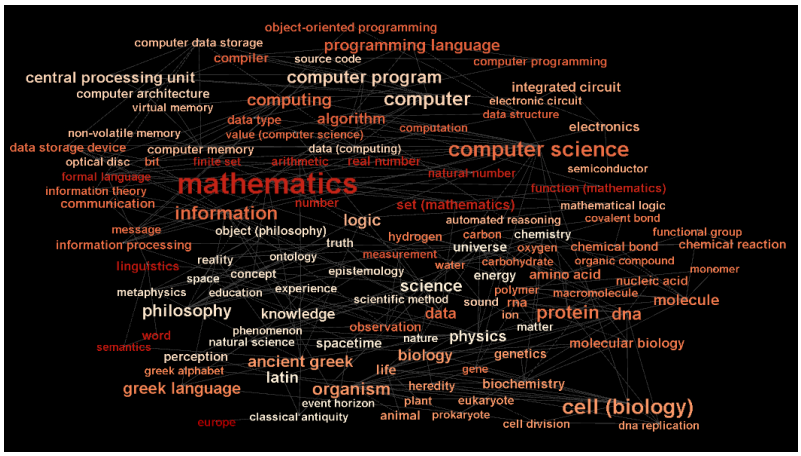


Figure 13.2: CN2 communities.

and entries that are related to the philosophy domain (*reality, concept, knowledge*, etc.) are in the white-coloured community. It can be observed that entries grouped into communities are more closely semantically related than entries from different communities. The results are similar for other networks; semantically related entries are grouped into communities much more than entries

that are not semantically related.

13.6 Conclusion

In this Chapter we present our initial attempt to study Wikipedia as a complex network. We extract parts of Wikipedia related to 5 chosen seed entries. We construct 10 different networks using two different principles of construction. Then we analyse the global structure of all networks. We show that all networks have similar properties: a high average clustering coefficient in comparison to the random networks, small distances, low density and community structure. From these global measures we may conclude that all 10 networks extracted form Wikipedia are small-world networks. These results are in line with previous studies of Wikipedia as a complex network.

Furthermore, we explore semantic relations in the constructed networks. We use network centrality measures to extract entries in the networks that are significant according to the network structure. Three centrality measures are employed for this task: degree centrality, betweenness centrality and closeness centrality. It can be observed that for level 2 networks centrality measures obtain good results (especially closeness centrality). Among top ten entries according to the closeness centrality are entries that are semantically related to the domain. This can be useful for modelling taxonomy or domain ontology. Furthermore, semantically related entries are grouped into communities more often that entries that are not semantically related.

These findings can be partially explained as a consequence of network construction rules employed in this experiment. However, these preliminary results suggest that Wikipedia is well organised and its structure can be captured and explored by a complex networks approach. In future work we plan to extract a broader section of Wikipedia and explore its potential as a knowledge network. We will study the domain knowledge extraction possibilities and perform the evaluation of the results.

14. Comparing Network Centrality Measures as Tools for Identifying Key Concepts in Complex Networks: a Case of Wikipedia

14.1 Abstract

Network centralities are amongst the most important measures for tracking and locating crucial nodes in a network. In this Chapter, we propose a general approach for identifying the most suitable centrality measure for detecting key concepts in a semantic or linguistic network. We experiment with seven network centrality measures (degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, current-flow betweenness centrality, current-flow closeness centrality and communicability centrality). For the purpose of evaluation, we compare the original Wikipedia hyperlink network with a constructed concept network. The obtained results indicate that all seven used measures have good potential for identifying key terms, and that degree centrality achieves the best score. A good score is also obtained for current-flow betweenness centrality and current-flow closeness centrality.

14.2 Introduction

The essential component of network science is a mathematical concept which we call a graph or a network. A graph, generally speaking, is represented as objects connected according to their relations. These objects are usually called vertices (nodes), and they are interconnected with edges (links). When we think of networks, we usually focus on representing some real-world relationships. Many objects of interest in the physical, biological, and social sciences can be represented as networks. Real-world networks are often complex networks which differ from regular or random networks in the fact that they exhibit some specific features like a community or hierarchical structure, giant components, a power law degree distribution, short average path lengths and high clustering coefficients [67]. Upon the construction of a network, we can analyze it utilizing various methods and metrics in order to extrapolate information pertinent to the network which are not immediately observable through its mere visualization. For instance, we may analyze a computer network in order to deduce how tolerant it is to attacks and will the vulnerability of certain nodes result in the loss of data flow. Another example is analyzing social networks to reason

about influencers [218] or to model knowledge flow through the network [219]. A prominent aspect of complex network analysis is the identification of important nodes in a network [220] which gives special interest to network centrality measures as indicators of which nodes have the crucial position in a network. Centrality measures may refer to the dominance of single nodes and are important in the construction of maximally efficient communication networks [221].

Furthermore, centrality measures indicate which nodes occupy important positions in the network . These measures were initially exploited in the domain of social sciences. The sociologist Freeman introduced betweenness-based centrality measures in [221]. Later on, Bonachich proposed the Eigenvector centrality measure [222]. These measures were later imported into other domains of complex networks like biological [223, 226] and infrastructure networks [227, 228]. Since then, many other centrality measures were proposed, specified for different tasks and ways of ranking nodes [224, 225, 229, 230].

In the domain of semantic and language networks, centrality measures have mainly been used for identification of keywords or key phrases [80, 87, 96, 99, 106, 231] and text summarization [85, 98, 104] .

The results of previous analyses of language networks motivated us to analyze centrality measures in the context of Wikipedia. We have already analyzed and compared the potential of different centrality measures for keyword extraction from texts [72, 201]. In [202] we proposed a new method for keyword extraction based on the selectivity measure. Wikipedia is interesting to study from different aspects . In [31], we analyzed networks of syllables constructed from texts found on Wikipedia. Furthermore, we experimented with the extraction of domain knowledge from Wikipedia [232]. In this work we describe a new approach for identifying key concepts in Wikipedia texts (entries) by means of seven network centrality measures: degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, current-flow betweenness centrality, current-flow closeness centrality and communicability centrality and rank them according to their performance in the evaluation procedure. Although centrality measures have been widely used for keyword extraction, to the best of our knowledge, current-flow betweenness centrality and current-flow closeness centrality were used for key concept identification for the first time. Moreover, the novelty of the proposed approach lies in the fact that it utilizes key concepts for construction of a concept network. More precisely, the presented algorithm identifies semantically related articles based on the keywords they share .

In the presented experiment, we treat foremost centrally positioned nodes as key concepts in a complex network constructed around Wikipedia's linked structure . The main goal of the presented experiment is to identify and explore which of the seven measures is suitable for the task of identifying central concepts in a semantic or linguistic network. Centrality measures are used in order to look at how centralities fare amongst themselves when considering the quality of Wikipedia's link structure contrasted with the semantic content found in the texts themselves. For the purpose of the analysis, comparison and evaluation of the set of centrality measures, we propose an approach based on the assumption that Wikipedia entries with a certain number of shared central concepts should be linked . Therefore, we construct a concept network in which Wikipedia entries are nodes, and a link between two entries is established if these two entries have a certain number of central concepts in common.

Next, we perform the evaluation procedure in which we measure the amount of overlap between the constructed concept network and a real network of hyperlinked Wikipedia entries. The overlap is measured in terms of the Jaccard index.

The remainder of the Chapter 14 is organized as follows. In the Section 14.3, we present related work about Wikipedia as a complex network and we give a short overview of the importance of centrality measures. In the Section 14.4, we give a definition of complex networks and provide equations and descriptions of all network centrality measures used in the presented experiment.

Moreover, we describe steps of the proposed approach for a three-layer network construction and evaluation of centrality measures. In the Section 14.5, we describe an experiment based on the proposed approach and in the fifth, we present the results of the conducted experiment. Finally, the Section 14.6 contains a conclusion and possible directions for future research.

14.3 Background and Related Work

14.3.1 Wikipedia as a Complex Network

Wikipedia is a free, online, collaborative, general knowledge encyclopedia . It was launched in 2001 and is currently available in 295 different languages. It is among the 10 most popular websites in the world, and its English language variant includes over 5.3 million unique entries (articles) [233]. Wikipedia is one of the largest open access compendiums of human knowledge and is updated daily by a workforce of over 134,711 regular volunteer editors [234]. As far as the validity and quality of Wikipedia entries are concerned, a 2005 study published in Nature showed that Wikipedia averaged 3.86 errors per entry. Contrasted with the 2.92 errors per entry average of the de facto standard, which is the Encyclopedia Britannica, Wikipedia proved its status as a valuable knowledge resource [233].

Wikipedia, as most encyclopedias, revolves around individual entries. As is typical for WWW documents, it is a hypertext wherein normal text is interspersed with hyperlinks pointing towards other related Wikipedia entries. Since an encyclopedia of this type strives to have its entries mutually well connected in order to facilitate the traversal of relevant topics, the number of hyperlinks is usually rather high. This connectedness of Wikipedia entries is the most basic principle following which complex networks are constructed from entries and their hyperlink structure . The model to construct a network relies on taking a starting entry as a seed node and then building edges according to the appearance of hyperlinks, each new hyperlinked entry being a new node within the network. Having a methodology for constructing networks out of knowledge embedded in Wikipedia's entries, we are able to extrapolate new knowledge pertaining to the chosen networks of concepts and Wikipedia at large.

Early attempts to quantify Wikipedia using complex networks analysis were focused only on the network structure of linked Wikipedia entries. In [216] Zlatić et al. present an analysis of Wikipedias in several languages as complex networks. They show that many network characteristics (degree distributions, growth, topology, reciprocity, clustering, assortativity, path lengths and triad significance profiles) are common to Wikipedias in different languages and show the existence of a unique growth process. The same authors studied Wikipedia growth based on information exchange in [217]. In [212], the authors presented an analysis of the statistical properties and growth of Wikipedia. Pemble and Bingol [27] have constructed two complex networks out of English and German Wikipedias and analyzed conceptual networks in different languages. Other research is focused on content and analyzes Wikipedia as a (domain) knowledge network.

Fang et al. [213] extract a specific domain knowledge network from Wikipedia (specifically, four domain networks on mathematics, physics, biology, and chemistry). They first present an efficient method to extract a specific domain knowledge network from Wikipedia. Furthermore, they carry out statistical analysis on four constructed knowledge networks. They show that MathWorld and Wikipedia Math share a similar internal structure. In [214], Masucci et al. extract the topology of the semantic space of Wikipedia entries. They find that the topology of the semantic space is scale-free in its connectivity distribution and displays small-world properties. They further measure semantic flow between different Wikipedia entries (represented as a directed complex network) and reveal the Scale-Free Architecture of the Semantic Space. In [235] authors construct four complex networks of different areas (Biology, Mathematics, Physics, and Medicine) based on cross-citations in the English version of Wikipedia. Entries are nodes, and the citations among

the entries correspond to edges. They analyze the clustering coefficient, topological structure, degree distribution, assortativity, betweenness centrality and average shortest path length . Their results indicate that analysis of the full Wikipedia network cannot predict the behavior of isolated categories since their properties can be very different from those observed in the full network.

Furthermore, there are certain attempts at link prediction on Wikipedia as a hyperlinked network. In [236], authors are dealing with the task of link prediction in the structure of hyperlinked document collections in Wikipedia . They propose a novel approach based on principal component analysis which relies only on hyperlinks, not on the textual content of entries. The conducted evaluation of the proposed approach shows that it improves the identification of the top missing links. Additionally, the proposed approach can be used to identify topics an entry misses to cover and to cluster entries semantically. In [237], authors explore statistical properties of links within Wikipedia. They show that algorithms based only on the hyperlink structure (not on topics) can predict new links. However, a topic-oriented PageRank algorithm can effectively identify topical links within existing entries. Based on these results, the authors propose a link prediction approach that combines structural requirements and topical relationships within Wikipedia.

14.3.2 The Role of Centrality Measures

The role of centrality measures is to identify the most important nodes in a network's architecture [238]. There are different definitions of centrality, depending on how we define a node's importance. Centrality measures are discriminative properties of the importance of a node in a graph and are directly related to its structure [75]. Therefore, centrality measures have the potential to extract key concepts from co-occurrence networks of texts. There are many studies in which various centrality measures are exploited for the task of keyword and keyphrase identification. The extensive related work on network centrality measures used for keyword extraction is reported in [72]. Here we discuss only some of the approaches relevant for this study.

Mihalcea and Tarau in [104] introduce a state-of-the-art TextRank algorithm (derived from PageRank) for keyword extraction. Boudin [80] compares various centrality measures for graph-based key phrase extraction. He shows that simple degree centrality obtains results comparable to the widely used TextRank algorithm; and that closeness centrality achieves the best results on short documents. Litvak and Last [99] test approaches based on the graph-based syntactic representation of text and web documents. They show that simple degree-based rankings from the first iteration of HITS already have satisfactory results. Lahiri et al. [96] extracted keywords and keyphrases from co-occurrence networks of words. They test eleven measures (degree, strength, neighborhood size, coreness, clustering coefficient, structural diversity index, page rank, HITS hub and authority score, betweenness, closeness and eigenvector centrality) and show that simple measures like degree and strength outperform coreness and betweenness which are computationally more expensive.

Obviously, various centrality measures can be used for the identification of key concepts. In this research, we adopt that assumption and aim to identify which measure is most suitable for identifying key concepts within Wikipedia texts. We carry out an evaluation based on the original Wikipedia hyperlink network. The performed evaluation is based on the fact that centrality measures play an important role in link prediction. This idea can be corroborated by the fact that preferential attachment is a well-known local similarity measure used predicting links on a local level. For example, in [191], authors develop a supervised learning approach to link prediction using a feature set of graph measures chosen to capture a wide range of topological structures. They include node centrality measures for link prediction.

To summarize, our approach assumes two things: first, that centrality measures can extract important key concepts as a set of top-rated nodes and second, that entries with a certain number of key concepts in common can be linked in the original Wikipedia hyperlink network.

14.4 Methodology

14.4.1 Complex Networks

A graph is an ordered pair $G = (V, E)$ where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. A graph is directed if edges have a direction associated with them. A graph is weighted if there is a weight function w that assigns value (a real number) to each edge. The number of nodes and edges in a graph is denoted as $N = |V|$ and $K = |E|$. A path in a graph is a sequence of edges which connects a sequence of nodes which are all distinct from one another. A shortest path d_{ij} between two nodes i and j is a path with the shortest length and it is called the distance between i and j.

14.4.2 Network Centrality Measures

In this Section, we provide explanations and equations for centrality measures used in our experiment.

Degree centrality of a node is determined according to (in- and out-degree in the case of directed networks) the number of nodes with which it is connected. When normalized by dividing it by the maximum possible degree $N-1$ we get the following equation:

$$C_d(v) = \frac{d(v)}{N-1}. \tag{14.1}$$

Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes, i.e. it measures how many time the node is on the network's shortest path. Nodes with high betweenness centrality may have considerable influence within a network by virtue of their control over information passing between other nodes. It differs from other centrality measures in principally not being a measure of how well-connected a node is. Instead, it measures how much a node falls between others or controls flows between others . Let σ_{jk} be the number of shortest paths from node j to node k and let $\sigma_{jk}(i)$ be the number of those paths that pass through node i. The normalized betweenness centrality of a node i is then given as:

$$C_b(v) = \frac{\Sigma_{u \neq v \neq t} \frac{\sigma_{ut}(v)}{\sigma_{ut}}}{(N-1)(N-2)}. \tag{14.2}$$

Closeness centrality is defined as the mean distance from a node to all other reachable nodes . In other words, it is the inverse of farness, i.e. the sum of the shortest paths between a node and all other nodes. So the closer a node is, the lesser its distance to all other nodes in a network. The normalized closeness centrality of a node i is then given by:

$$C_c(v) = \frac{N-1}{\Sigma_{v \neq u} d_{vu}}. \tag{14.3}$$

Eigenvector centrality can be thought of as an upgrade of standard degree centrality . Degree centrality measures only the amount of connections a node has but disregards towards which nodes these connections are established. Eigenvector centrality modifies this approach by giving a higher centrality score to those connections which are made towards those nodes which are themselves central. Thus, it measures influence within a network. A node's eigenvector centrality has the useful property that it can be large either because it has many neighbors or because it has important neighbors (or both). Also, the centrality C_{EV} of node v is proportional to the sum of the centralities of its neighbors. For the node v and constant λ it is defined:

$$C_{EV}(v) = \frac{1}{\lambda} \Sigma_{u \in N(v)} C_{EV}(u). \tag{14.4}$$

Current-flow centralities are variations on the classical betweenness and closeness centralities originally proposed in [229] . These measures take into account that information spread is calculated via the assumption that it spreads as efficiently as an electrical current (current-flow). Each link is given an arbitrary orientation, so $\overrightarrow{e}$ denotes the directed link corresponding to the orientation of $e \in E$. Furthermore, the authors define the throughput of a node $v \in V$ for a given supply b and $x(\overrightarrow{e})$ defined as an electrical current vector (for more details see [229]):

$$\tau(v) = \frac{1}{2}(-|b(v)| + \Sigma_{e:v\in v}|x(\overrightarrow{e})|). \tag{14.5}$$

Finally, **current-flow betweenness centrality** is defined as follows :

$$C_{CFB}(v) = \frac{\Sigma_{s,t\in V}\tau_{st}(v)}{(N-1)(N-2)}, \tag{14.6}$$

where τ_{st} denotes the throughput in case of an st-current.

Current-flow closeness centrality is defined as :

$$C_{CFC}(v) = \frac{N-1}{\Sigma_{t\neq s}p_{st}(s) - p_{st}(t)}, \tag{14.7}$$

where $p_{st}(s) - p_{st}(t)$ corresponds to the effective resistance, which can be interpreted as an alternative measure of distance between s and t.

Communicability centrality is another measure closely tied to betweenness centrality [224, 225] . Instead of considering just paths passing through nodes in a network, communicability centrality introduces scaling so that not all paths are seen to be of equal worth, longer paths obviously having a lower value. As such, it measures how easy it is to pass messages between nodes in a network. We can interpret the local communicability of a node as a measure of how well connected it is. Global communicability of the entire network can, for instance, help us discover bottlenecks. Communicability between two nodes v and u can be calculated as the weighted sum $com(v,u)$ of all walks between nodes v and u. Then the total communicability of a node v is given as:

$$C_{CC}(v) = \Sigma_{u\in N}com(v,u). \tag{14.8}$$

14.4.3 The Proposed Approach

Here we describe an approach for comparing network centrality measures as tools for identifying concepts in complex networks. The main idea is a three-layer network construction in which networks on the third layer show which entries are semantically close and share key concepts. For the purpose of evaluation of this assumption, the last step of our experiment compares networks of the third layer with the original network of hyperlinks on the first layer. The proposed three layers of networks based on Wikipedia are:

- The **first layer**, L1 is the network of hyperlinks. This is the original network of Wikipedia hyperlinks which serves as a referential model in the evaluation step.
- The **second layer**, L2 is a set of co-occurrence networks based on texts extracted from each of Wikipedia's entries. In these networks nodes are words, and two nodes are connected if they co-occurred as neighboring words in the same sentence in the text. This is just an auxiliary network which is used for extracting key concepts from an entry. Key concepts are then identified using different network centrality measures.

- The **third layer**, L3 is a concept network built upon the second layer by connecting two entries if they share a certain number of key concepts (for different thresholds and different centrality measures).

The details of the entire experiment are described as follows.

For the construction of **the first layer**, it is necessary to construct a hyperlink network. In general, this network may contain the entirety of Wikipedia. However, due to its large scale, we introduce certain limitations. Firstly, we choose one seed entry as a starting point from which our network of hyperlinks will be constructed. Secondly, we chose a limited number of hyperlinks from the seed entry to collect new entries. Thirdly, we limit the number of times (the depth of the hyperlink network) that we would repeat the whole collection procedure. More precisely, we introduce three limitation parameters: the seed entry (SE), the number of collected hyperlinks (NL) and the hyperlink network depth (ND). The first layer is then a hyperlink network - a subset of the whole Wikipedia hyperlink network, $L_1 = G_H = (V_H, E_H)$. Every hyperlink network is originally a directed network. However, for the purposes of comparison and evaluation, the constructed network will be observed as undirected.

For the construction of **the second layer**, it is necessary to extract the text from each entry collected in the previous step. After that, texts should be preprocessed and prepared for the construction of co-occurrence networks. The preprocessing of texts includes transformation into lower caps, the removal of punctuation and stop words, and lemmatization. For each text, a co-occurrence network is constructed. A co-occurrence network is a network created by getting a Wikipedia entry's text and connecting the nodes, each node being a single word, in such a way that words occurring immediately after one another are connected. The result of this step is a second layer which is a set of co-occurrence networks, $L_2 = G_1 = (V_1, E_1), ..., G_k(V_k, E_k)$. The number of networks is equal to the number of nodes in the hyperlink network.

Finally, the construction of **the third layer** network is based on the second layer. The nodes are entries and two nodes (entries) are connected if entries share a certain number of key concepts. Here again we need to define certain parameters in order to specify the key concepts and their number. Key concepts can be identified by choosing a network centrality measure. Therefore, first we need to specify a centrality measure (CM) that will be used for the construction. The result of applying the centrality measure to one network (entry) is a ranking list of all the nodes in the network. Nodes represent words, and highly ranked words can be assumed to be key concepts in the entry. Then, we need to determine how many words from the ranked list will be used as key concepts (NKC). Lastly, we need to set a threshold (t). The threshold is the number of the minimum key concepts that two entries should have in common in order to be deemed related and connected with an edge. The result is a new network, $L_3 = G_C = (V_C, E_C)$. We call it a concept network because it represents how Wikipedia concepts are related according to the chosen centrality measure. The concept network has the same set of nodes as the original hyperlink network ($V_C = V_H$), but a different set of edges. This network is observed as a weighted network where the weight represents the number of shared key concepts. The weights are ignored for the purposes of evaluation.

The described procedure can be summarized as an algorithm performed in six main steps outlined in the algorithm for three-layer construction.

Now it is possible to compare the concept network, G_C with the hyperlink network, G_H. The comparison is performed via the Jaccard index, also known as the Jaccard overlap or the Jaccard similarity coefficient for comparing sets. It is defined by the following equation:

$$JI(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{14.9}$$

According to the Jaccard index, we are focused only on that part of the concept network that is the subset of the hyperlink network (as it is shown in Figure 14.1), but it is also possible to

ALGORITHM three-layer construction

INPUT: SE, NL, ND, CM, NKC, t

OUTPUT: three-layer of networks

1: Creation of a hyperlink network (N_H) using a seed entry of choice (SE), by collecting first NL hyperlinks and repeating the procedure ND times.

2: Extraction of a complete entry text for every node in the previously constructed network. Text preprocessing by means of: Transformation of each text into lower caps. Removal of punctuations from each text. Removal of stop words and lemmatization of each text.

3: Creation of a set of co-occurrence networks from texts $\{G_1, \ldots, G_k\}$.

4: Extraction of top NKC key concepts from each network (text) according to the chosen centrality measure (CM).

5: Creation of a concept network (N_C) taking into account only t overlapping entries.

6: RETURN: Set of three layers of networks

$$\begin{Bmatrix} L_1 = \{G_H = (V_H, E_H)\}, \\ L_2 = \{G_1 = (V_1, E_1), \ldots, G_k = (V_k, E_k)\}, \\ L_3 = \{G_C = (V_C, E_C)\} \end{Bmatrix}$$

analyze the whole concept network. In this case, the observed part of the concept network is an overlapping network (a subset network in Figure 14.2), $G_{ovp} = (V_{ovp}, E_{ovp})$ where $V_{ovp} = V_H \cap V_C$ and $E_{ovp} = E_H \cap E_C$. In terms of our experiment, we need to compare the hyperlink network's set of edges with that of the concept network.

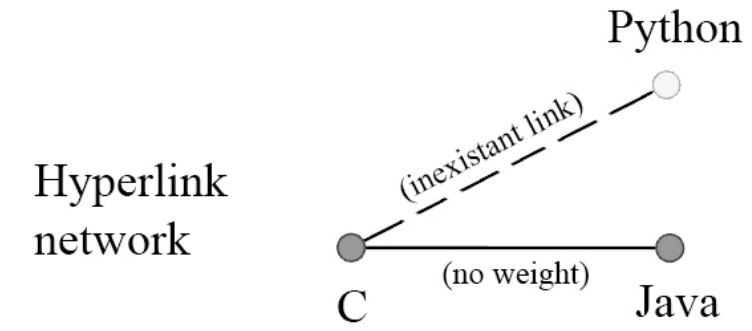
$$JI(E_H, E_C) = \frac{|E_H \cap E_C|}{|E_H \cup E_C|}. \tag{14.10}$$

The described procedure can be performed repeatedly with different parameters and various centrality measures to gain better insight into which centrality measure is the most appropriate for extracting key concepts from Wikipedia entries. Furthermore, the same procedure can be exploited with the aim of proposing possible missing links in the original hyperlink network. Missing links can be proposed from the set of edges that exist in the concept network and do not exist in the hyperlink network. In the presented experiment, we are focused only on the first part of the task and in the following Section we present a case study in which we compare seven centrality measures.

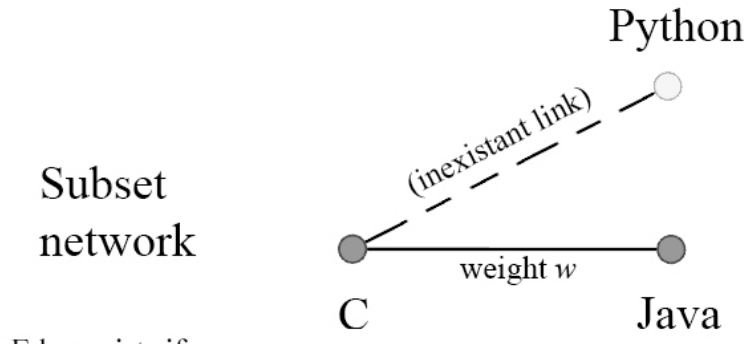
14.5 Experiment Description: Datasets and Network Construction

For the purpose of the presented experiment, the network of choice has a seed entry "Programming language" (SE = *"Programming language"*), the number of hyperlinks is set to 20 ($NL = 20$) and the hyperlink network depth is set to 2 ($ND = 2$). Starting with a chosen seed entry, we store all the hyperlinks to related entries from the seed entry's text (depth 1) and proceed to extract the hyperlinks from all the entry pages taken from the original entry (depth 2).

Therefore, the first task is the implementation of a web scraping program which extracts hyperlinks from a Wikipedia entry's text. The hyperlinks are extracted using a Python package for HTML parsing called Beautiful Soup [239] which parses the HTML structure of a given HTML document into a parse tree. By navigating the tree one can locate the tag ID which corresponds to entry content ("mw-content-text") and proceed to extract the hyperlinks which themselves are



Seed entry: "*programming language*"
Depth: level 2
Edge exists if hyperlink exists.



Edge exists if:
1) Hyperlink exists in the hyperlink network.
AND
2) $w \geq$ threshold t

Figure 14.1: The original hyperlink network constructed on the first layer (above) and a subset of the concept network on the third layer (below). Both networks have the same number of nodes. In the concept network, the edge between two nodes exists if these two nodes (entries) have a certain number of common key concepts ($\geq t$). The subset network is a part of the concept network which intersects with the hyperlink network.

found within paragraph (<p>) tags and finally inside link (<a>) tags in that section of the page. The network is stored as an edge list. In such a network, each entry's title represents a node and it is connected to other entries hyperlinked in its text, again represented as network nodes. The hyperlink network $G_H = (V_H, E_H)$ constructed from the chosen seed entry has 302 nodes and 356 edges.

Then we construct a set of 302 co-occurrence networks, $L_2 = N_1 = (V_1, E_1), ..., N_{302} = (V_{302}, E_{302n})$. Each network is based on one Wikipedia entry text . For each text, a co-occurrence network is constructed according to the rule that all the words are nodes and two nodes (words) are connected if and only if these two words are neighboring words in the same sentence. Before network construction, we perform text preprocessing. Lemmatization was done by using the NLTK Python

toolkit (Natural Language Toolkit), [33] and the included Wordnet lemmatizer. The list of stop words that we used in order to prepare the texts for the creation of co-occurrence networks was borrowed from Wikiminer [240] and later expanded on our own with suitable stop words that were found missing from the original list. The removal of stop words and punctuation, and the creation of co-occurrence networks was accomplished by using the LaNCoA toolkit (Language Networks Construction and Analysis) , [8]. Additionally, we used Python and the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [161].

Next, we construct various concept networks, $G_C = (V_C, E_C)$ with different parameters. The chosen centrality measures were: degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, current-flow betweenness centrality, current-flow closeness centrality and communicability centrality. After experimenting with different values, we set the NKC value to 20, i.e. we choose 20 top key concepts ranked by the chosen centrality measure. By setting the NKC value to 30, we get a more densely connected concept network, while we get the opposite effect by setting the NKC to lower values. That helped us conclude concluded that value 20 is the best for the NKC parameter in the case of Wikipedia. Then we experimented with three thresholds: $t = 1, t = 3, t = 5$ and realized we have an opposite situation with t compared to NKC.

The creation of 3 new networks for each centrality measure resulted in 21 networks. All 21 networks were then compared with the original hyperlink network via the Jaccard index. The detailed procedure of the preformed experiment is depicted in Figure 14.2.

14.6 Results

In this Section, we present the results of the evaluation procedure for seven centrality measures used to identify key concepts of Wikipedia entries and compare them to determine which one gives the top performing result. Guided by the notion that two entries are semantically related and linked in the original hyperlink network if they share a certain number of key concepts, we provide a comparison of centrality measures based on the original hyperlink network as the referential model.

Each of the Tables 14.1, 14.2 and 14.3 (each table for one threshold) serves to show the comparison between the original hyperlink network and the 21 concept networks. Each row in the table represents one centrality measure. The first two columns merely specify the basic metrics (the number of overlapping nodes, $N_{ovp} = |V_H \cap V_C|$ and the number of intersecting edges $K_{ovp} = |E_H \cap E_C|$). The third column specifies the Jaccard index (JI) which is a measure of similarity between the hyperlink network and the concept network at hand. According to the equation (10), it is calculated by dividing the number of links that the two networks have in common ($Kovp$) with the total number of links in the hyperlink network ($K_H = 356$). The last column shows the centrality measure rank according to the Jaccard index.

Centrality measure	N_{ovp}	K_{ovp}	JI	$Rank$
Closeness (C_c)	265	314	0.8792	6.
Betweenness (C_b)	274	323	0.9044	4.
Eigenvector (C_e)	264	307	0.8595	7.
Degree (C_d)	**283**	**333**	**0.9325**	**1.**
Current-flow betweenness (C_{cfb})	278	328	0.9185	2.
Current-flow closeness (C_{cfc})	275	325	0.9101	3.
Communicability (C_{com})	273	322	0.8988	5.

Table 14.1: Performance of centrality measures with threshold $t = 1$.

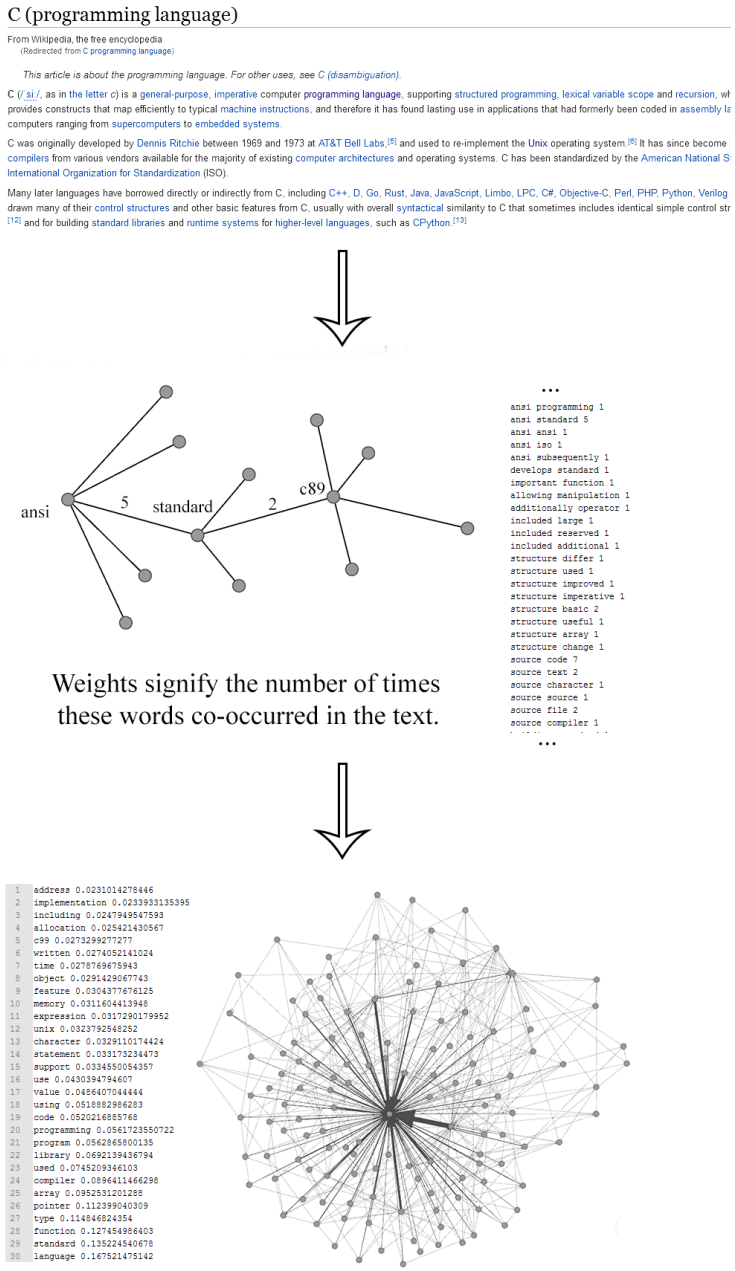


Figure 14.2: Details of the entire experiment for the chosen seed entry *"Programming language"*.

In Figure 14.3 we plot the overall performance for all seven measures and the three different thresholds shown in blue $(t = 1)$, red $(t = 2)$ and green $(t = 3)$. As expected, the higher the threshold needed to establish a link between nodes (concepts), the lower the similarity between the networks.

Although overall results show that there are no significant differences among the seven measures, degree centrality noticeably performs best for all thresholds, while eigenvector centrality exposes the lowest potential in this task. These results are in line with results presented in [80, 87, 96] which proved that degree centrality is a suitable network measure for extracting key terms from texts, regardless of the used threshold value.

The current-flow betweenness and current-flow closeness centralities evaluate right underneath

Centrality measure	N_{ovp}	K_{ovp}	JI	$Rank$
Closeness (C_c)	153	170	0.4747	6.
Betweenness (C_b)	199	224	0.6264	4.
Eigenvector (C_e)	124	142	0.3960	7.
Degree (C_d)	**202**	**235**	**0.6573**	**1.**
Current-flow betweenness (C_{cfb})	200	231	0.6460	2.
Current-flow closeness (C_{cfc})	194	226	0.6320	3.
Communicability (C_{com})	182	217	0.6067	5.

Table 14.2: Performance of centrality measures with threshold $t = 3$.

Centrality measure	N_{ovp}	K_{ovp}	JI	$Rank$
Closeness (C_c)	68	64	0.1797	6.
Betweenness (C_b)	111	116	0.3230	4.
Eigenvector (C_e)	61	60	0.1657	7.
Degree (C_d)	**122**	**132**	**0.3679**	**1.**
Current-flow betweenness (C_{cfb})	120	131	0.3651	2.
Current-flow closeness (C_{cfc})	112	119	0.3314	3.
Communicability (C_{com})	95	96	0.2668	5.

Table 14.3: Performance of centrality measures with threshold $t = 5$.

it regardless of the threshold value. Closeness and eigenvector measures are underperforming since they are evaluated as lowest performing measures, regardless of the threshold.

This work is the first attempt to test current-flow betweenness centrality, current-flow closeness centrality and communicability centrality in the task of keyword extraction. Here we report that all three measures show good results in the task of identifying key concepts and current-flow betweenness centrality almost yields the best results.

14.7 Conclusion

In this study, we analyze the potential of network centrality measures for identifying key concepts in Wikipedia texts. The presented experiment is built upon two assumptions about networks: (1) network centrality measures can identify key concepts (words) in co-occurrence networks of texts; (2) entries with a certain number of mutual concepts are more likely to be connected and linked.

Obtained results confirm that network centrality measures have much potential for the extraction of key terms in general. In this experiment, some centrality measures perform better (degree centrality, current-flow betweenness centrality and current-flow closeness centrality) than others (eigenvector centrality, closeness centrality, communicability centrality and betweenness centrality).

This is the first time that current-flow betweenness centrality, current-flow closeness centrality and communicability measures were applied in the task of the identification of key terms. In this experiment, current-flow betweenness centrality and current-flow closeness centrality outperform standard betweenness and closeness centralities. This may be due to the fact that current-flow closeness centrality is equal to information centrality [229]. In contrast to common shortest-path-based centrality measures, information centrality takes into account all parallel paths. The same holds true for current-flow betweenness centrality. It seems that in co-occurrence language networks not only shortest paths are important. That makes sense since sentences may either be short or long and key terms are positioned on different paths.

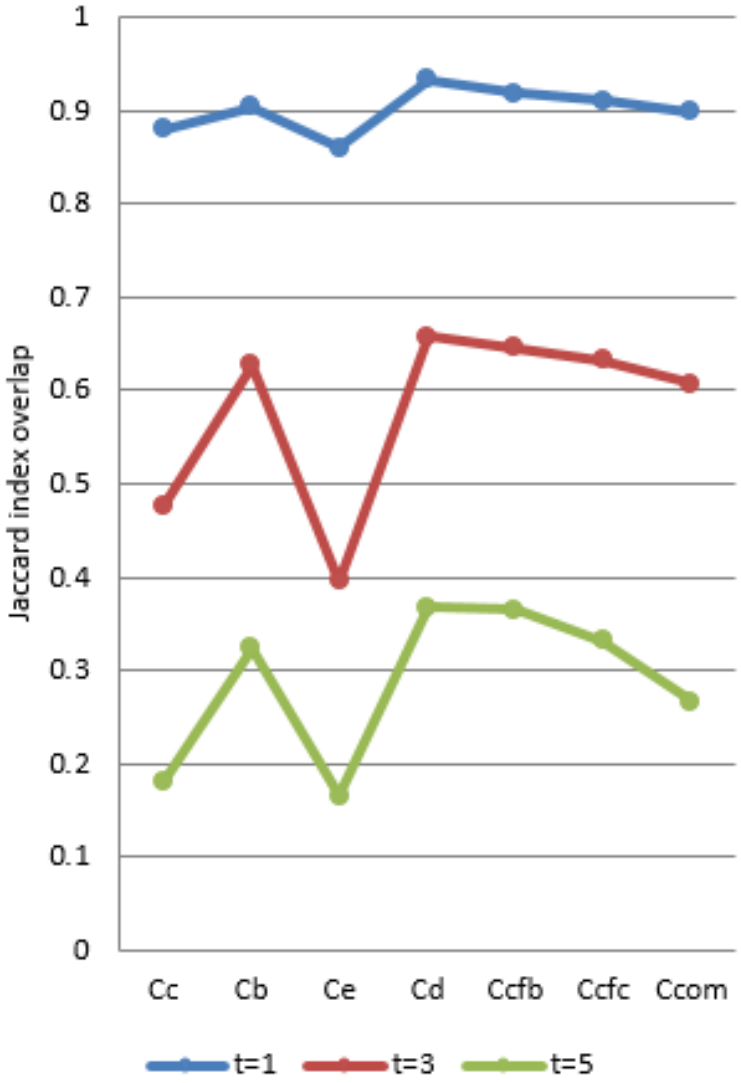


Figure 14.3: The performance of seven centrality measures combined with three thresholds ($t = 1; t = 3; t = 5$).

Another novelty of the described approach is that it proposes a particular evaluation procedure which is based on the underlying semantic relatedness of the concepts .

Overall, the two underlying contributions of this research are: (1) comparison of network centrality measures for identifying key concepts in the context of Wikipedia; (2) a specific evaluation procedure based on the semantic. Note that this evaluation procedure is appropriate only in the case of Wikipedia and similar networks.

There are two limitations of this experiment. Firstly, we did not include all existing measures in the experiment. We selected those measures which are reported to perform well with texts and three new measures which were not tested on texts yet. Secondly, we made the experiment with only one seed entry. In the future, we plan to extend the experiment by using more seed entries and more centrality measures, e.g. extensions of current-flow betweeness centrality defined in [241].

Still, it seems that all tested measures perform reasonably well for lower thresholds, while the results are more differentiated for higher thresholds. According to the second assumption mentioned above, the presented method could also be applicable to the problem of identification of

missing links. Hence, we plan to test the potential and performance of centrality measures for the task of link prediction.

III Multilayered Language Model

15. Towards a Formal Model of Language Networks

15.1 Abstract

Multilayer networks and related concepts have been used for the description and analysis of various complex systems in many fields, such as for example biological, physical, social and information sciences. In this Chapter we propose a formal model for language representation - Multilayer Language Network (MLN) which is based on multilayer network formalism. This work presents the first steps towards a universal formal model suitable for representation, analysis and comparison of languages both in their entirety as well as in their various characteristics and complexity. The goal of this research is to define a universal formal model suitable for representation, analysis and comparison of languages, considering various language levels (subsystems) and various language characteristics. So far diagnostic for language networks has been reported for isolated subsystems. MLN model has the potential to extent from isolated to integral diagnostics, enabling better insights in mutual interactions of language subsystems. Here we discuss initial steps in this direction. Furthermore, we present MLN model for English and Croatian language, considering word, syllable and grapheme language subsystems and various construction principles. For the analysis we apply standard network diagnostics and present obtained results.

15.2 Introduction

The complex networks theory has been recognized as a particularly powerful framework for studying phenomena from gene/protein or social interactions, to technological and infrastructure systems [67, 242]. This generated a swift development of various fields and opened new research avenues [67]. One of them is network linguistics [6, 7, 11, 22], which contributed significant results, ranging from new models of language evolution [7, 22], to the quantitative analysis of written novels [11]. Progress was also made in the development of tools for text analysis [243], and mechanisms for automatic detection of polysemy [244]. Particular emphasis is constantly given to semantic networks, whose structure and dynamics are being examined in detail [3].

Various aspects of natural language can be represented as complex networks [67], whose nodes depict linguistic units (e.g. words), while edges model their morphosyntactic, semantic and/or

pragmatic interactions [6, 22, 210]. This refers to language analysis through varying linguistic levels (syntactic, semantic, phonetic [22, 210]), the examination of language evolution [3], or the modeling of language acquisition [37]. A specific interest lies in language technologies, aiming at developing software able to consistently carry out a desired analysis of a given text: assess the quality of a summary, extract text context, key phrases and keywords extraction, disambiguate word senses, estimate the translation and determine subjectivity [245]. In short, the complexity of language as a natural evolving system is mirrored by the structural complexity of the corresponding network model.

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 [3, 6, 7, 11, 22, 245]. Small-world networks [20] have a small average shortest path length and a large clustering coefficient and scale-free networks [43] have power-law degree distribution.

In the era of "big data" beside of the explosive growth of data we are also witnessing the swift advances in the theoretical models of multilayer networks, suitable to consistently model different data sources in the same framework. However, the field of complex networks has shifted from the analysis of isolated network (capturing and modelling one aspect of the examined system) toward the analysis of the family of complex networks simultaneously modelling different phenomena (aspects) of examined system, or simultaneously modelling interactions and relationships among different subsystems.

This pursuit opened a variety of different theoretical models: multilayer networks [246, 261], multidimensional networks [247], multiplex networks [248, 249], interdependent networks [250] and networks of networks [251]. A thorough discussion that compares, contrasts and translates between theoretical notions of multilayer, multiplex, interdependent networks and networks of networks is given in [168].

Multilayer network approach has been addressed in study of the international trade analysis [253], social interactions in the massive on-line game [254], web-search queries [247], in transport and infrastructure [246, 248] and for examining the brain function [255]. However, although multilayer networks fit the language levels in a natural way, there have been no reports on multilayer language networks. So far there have been only efforts to model isolated phenomena of various language subsystems (e.g. co-occurrence [6, 11]) and examine their unique function through complex networks, failing to explain mechanism of their mutual interaction or interplay.

This Chapter presents the first steps towards a universal formal model suitable for representation, analysis and comparison of languages both in their entirety as well as in their various characteristics and complexity. Such a model would be more general and expressive then existing approaches [6, 22, 37, 243, 244]. Inspired by [168], we base our approach on multilayer networks. To the best of our knowledge this is the first work that models languages by means of multilayer networks.

The Chapter is organized as follows. Section 15.3 introduces the formal multilayer network model for languages. In Section 15.4 we focus on some diagnostics of the model and present some initial experiments and results. We conclude in Section 15.5 by pointing to future work.

15.3 Formal Model

This Chapter introduces a formal model for languages based on graphs or networks. For simplicity, we often interchange the terminology of a graph and a network. We aim to design such a model that is universal in the sense that is suitable for the representation of all languages, for both written and spoken forms, as well as for the comparison of various languages, and likewise suitable for the linguistic analysis of any given language characteristics.

Kivelä at al.in [168] review and unify the terminology of existing concepts for multilayer network structure and similar network structures from the literature. In order to relate to existing

research, methodology and diagnostics we have tried to design our model as close as possible to the general framework and notions given therein. On the other hand, given the specifics of what is modeled by the formalism, i.e. languages in several of their features, we have somewhat modified that framework and terminology.

Multilayer Language Network Model

A ***Multilayer Language Network*** (MLN) M is a quintuple $M = (V_M, E_M, V, L, C)$ where

- V is a nonempty set whose elements are called ***nodes*** ;
- C is a nonempty set of ***perspective elements*** ;
- L is a set of ***perspects*** where $\{L_0, L_1, L_2\}$ is a partition of C. Perspect L_0 is the ***language perspect***, L_1 is the ***hierarchy perspect*** and L_2 is the ***construction perspect***
- For perspect $L_1 = \{g_1, \ldots, g_k\}$ sequence $g_1, \ldots, g_k$ is the subsequence of the following sequence

$$discourse, sentence, phrase, syntagm, word, morphem, syllable, phoneme, grapheme \tag{15.1}$$

  called ***hierarchy***, which is denoted by $h_1, \ldots, h_9$ in short ;
- an element of the set $L_0 \times L_1 \times L_2$ is called a ***layer*** ;
- $V_M \subseteq V \times L_0 \times L_1 \times L_2$ is the set whose elements are called *MLN-**nodes*** ;
- $E_M \subseteq V_M \times V_M$ is the set of ***edges***.

An example of a multilayer language network model is given in Figure 15.1 .

Underlying graph of an *MLN* is defined naturally, i.e. considering all *MLN-nodes* as its nodes and E_M as the set of its edges. *Underlying graph* of an *MLN* model $M = (V_M, E_M, V, L, C)$ is the graph $G_M = (V_M, E_M)$.

For an example of an *underlying graph*, see Figure 15.2 containing the *underlying graph* corresponding to the *MNL* model depicted in Figure 15.1 . Each node in the *underlying graph* represents an *MLN-node*, e.g. $(a, word, syntax)$, $(b, word, syntax)$, $\ldots$, $(i, grapheme, shuffle)$ etc.

As custom in the graph theory, *edges* in an *MLN* may be directed or undirected, weighted or unweighted. Consequently, we differentiate between ***directed*** *MLN*s, ***undirected*** *MLN*s, ***weighted*** *MLN*s, and ***unweighted*** *MLN*s on the basis of the corresponding *underlying graph*. For example, the *MLN* model presented in Figure 15.1 is directed and unweighted.

The set of *nodes* contains all the elements under consideration, that is linguistic units such as sentences, words, syllables etc. that appear in the text that is modeled by the given *MLN*.

Language perspect denotes the type of the language or the particular languages under consideration, e.g. English and Croatian . *Construction perspect* reflects approaches to the analysis of language structure, such as analysis of the syntax of a given text, and the analysis of the same text but with randomized word order . From the linguistic point of view that arises from the specifics of the model *hierarchy perspect* is the most essential of perspects . It represents levels of language denoted by the hierarchy sequence (15.1). Different *MLN*-s may focus on some of the language levels while other levels may be out of the scope, which is reflected in the *perspective elements* of the *hierarhcy* that the *hierarchy perspect* contains.

A *layer* is specified by a single *perspective element* from each of the *perspects* . Moreover, *perspects* are studied in all combinations This means that *layers* specify all possible perspectives or views on a language. For example, some of the *layers* in the model shown in Figure 15.1 are

$$(croatian, word, syntax) \qquad \text{and} \qquad (croatian, syllable, shuffle) \ .$$

Notice that by the above definition, depending on the set V_M, some *layers* may be empty, with no *MLN-nodes* on them. Such empty *layers* represent combinations of *perspective elements* that are not important or not studied in that particular language analysis.

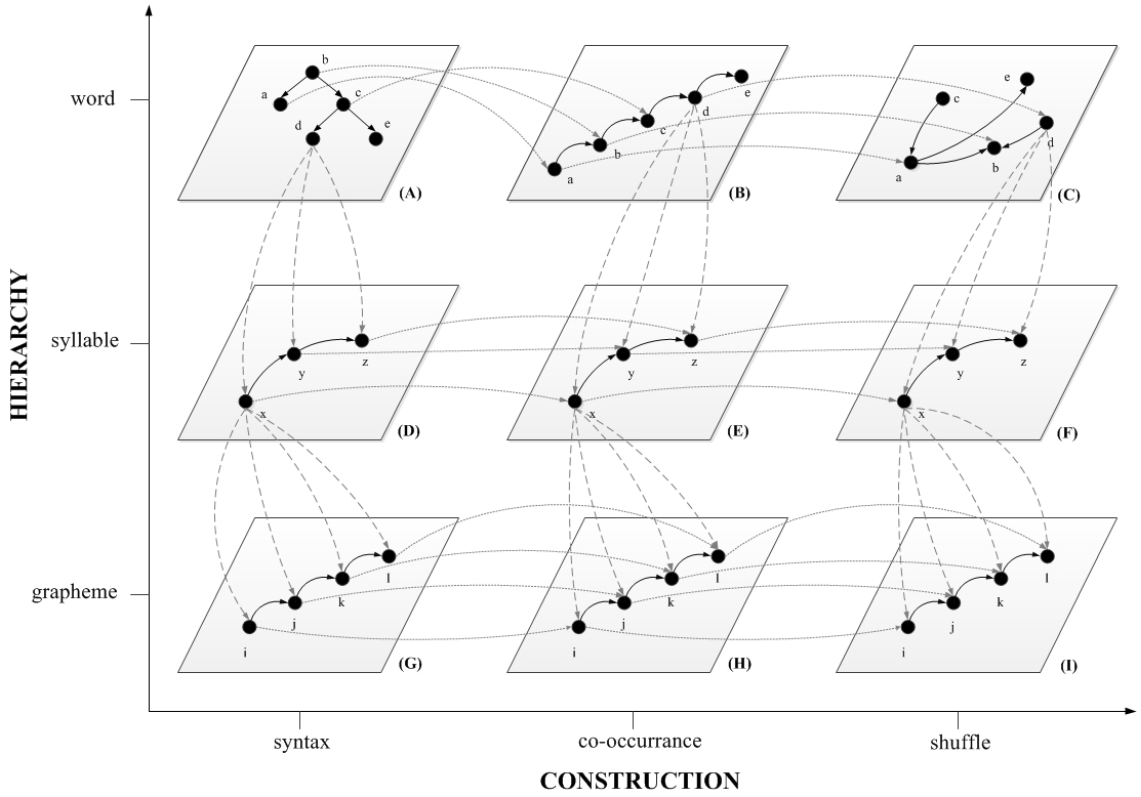


Figure 15.1: Example of an directed unweighted *MLN* model. The model consists of the following *perspects*: $L_0 = \{croatian\}$, $L_1 = \{word, syllable, grapheme\}$ and $L_2 = \{syntax,\ co\text{-}occurrence,\ shuffle\ \}$. The *language perspect* is omitted for simplicity. The remaining submodel contains the *hierarchy* and *construction perspect*. There are 9 *layers* denoted by $(A)-(I)$. *Intralayer edges* are represented by solid arrows, while the *interlayer edges* are dotted.

MLN-nodes are copies of *nodes* placed on different *layers*. For an *MLN-node* (a,l), where $a \in V$, $l \in L_0 \times L_1 \times L_2$, we say that the *node a* appears on *layer l*.

Informally, we may think of an *MLN* model as a graph or a network with some additional structure. In other words, *MLN* model and its *underlying graph* may be interchanged. Then *MLN-nodes* are "nodes" of that graph and E_M is the set of its "edges" . For simplicity, we will sometimes say *node* for an *MLN-node* when the meaning is clear from the context.

Edges in an *MLN* may be defined between any of the *MLN-nodes*. The set of *edges* in an *MLN* can, therefore, be partitioned into ***intralayer edges*** and ***interlayer edges***. An *intralayer edge* connects two *MLN-nodes* from the same *layer*, while an *interlayer edge* is an *edge* between two *MLN-nodes* belonging to different *layers*.

In case that some language analysis does not require some *perspect*, that *perspect* may be omitted from the model. In the same way new perspects can be introduced to the model, allowing analysis of some other phenomena of interest. Submodels may for example be used in the analysis of a single language. Then the *language perspect* would consist only of one language, e.g. Croatian in the example model depicted in Figure 15.1 , and can be omitted for simplicity. Such a submodel would only contain the remaining two *perspects*, *hierarchy* and *construction perspect*.

Kivelä et al in [168] review and classify multilayer networks based on the types of constraints imposed on the network. We, however, put no constraints on the model, both with respect to *nodes* as well as *edges*. More precisely, we allow *edges* between arbitrary *MLN-nodes* in the system but

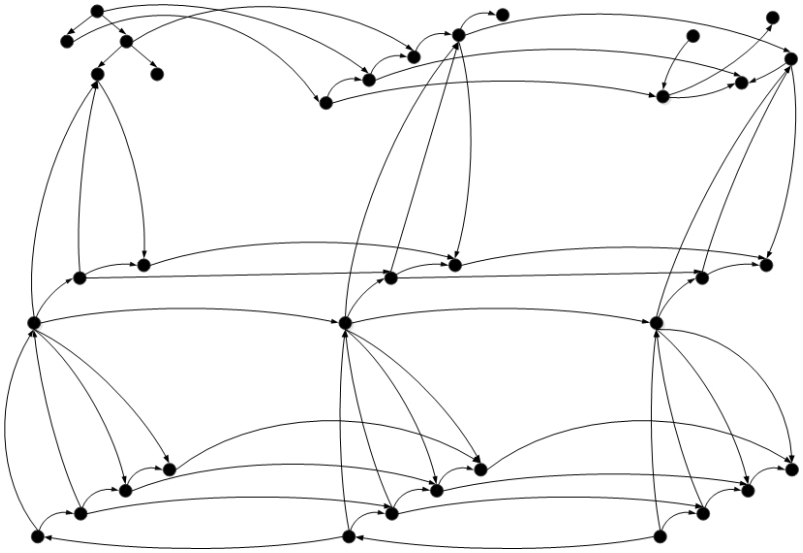


Figure 15.2: *Underlying graph* of the *MLN* model depicted in Figure 15.1.

do not impose *edges* between some *MLN-nodes* either. Also, a *node* does not need to appear on every *layer*. This is different from the multiplex structure [168, 248, 249] where traditionally all nodes appear on every layer, i.e. all nodes are shared between all layers. Additionally, in a multiplex there are edges between each node and its counterparts in different layers, that is, adjacency of a node with itself across multiple layers is explicit . At the same time, in a multiplex there is no adjacency of a node with other nodes from different layers. This is usually called diagonal coupling and categorical coupling in the literature [168].

One of the characteristics of the *MLN* model is that *hierarchy* imposes order of linguistic levels. The structure of language subsystems is preserved and modeled through *hiearchy perspect*. This is similar to ordinal couplings [168, 256], in which layers are ordered and nodes are adjacent only to their counterparts in consecutive ("adjacent") layers.

Another difference to the model presented in [168] is that we allow self-edges, that is an *MLN-node* being adjacent to itself. These *edges* would for example represent neighbouring words such as "bla bla bla" or *edges* between syllables in the same word, e.g. in "banana".

Although our model can represent structures such as multiplex in principle, we do not impose nor disallow adjacency. This approach allows wide and universal analysis and comparison of different linguistic phenomena.

15.3.1 Interpretation of *MLN*

MLN model has several features. It allows universal representation of languages, their analysis in an unifed framework, as well as comparison of various language phenomena .

For example, an *MLN* can model a particular Croatian novel and its linguistic units at chosen levels, e.g. all of the words, syllables and graphemes that appear in the novel. Besides original text of the novel, one could consider the same text but, for instance, with randomised word order on the sentence level, or consider the syntax dependencies of words in a sentence.

The model would have the following *perspects*:

$$L_0 = \{Croatian\},\ L_1 = \{word, syllable, grapheme\} \quad \text{and} \quad L_2 = \{syntax, co\text{-}occurrence,\ shuffle\}\ .$$

For simplicity, we can omit the first *perspect*. We would then consider the *MNL* submodel with 2 *perspects*, the *hierarchy* and *construction perspect*, as is the submodel shown in Figure 15.1 .

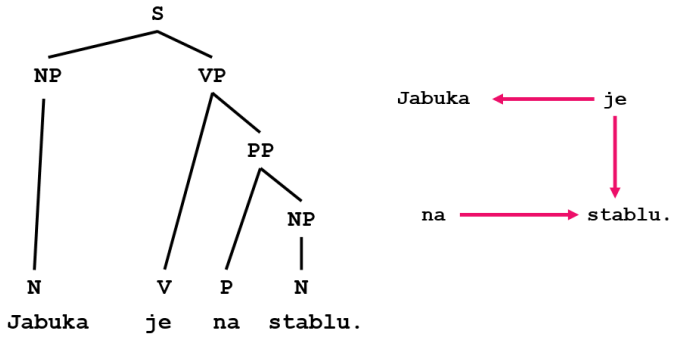


Figure 15.3: An example of the syntax depencency parsing for the sentence "*Jabuka je na stablu.*". This translates to "*The apple is on the tree.*" from Croatian. Syntax depencency tree for that sentence (as per [257]) is presented to the left. Corresponding graph is given to the right.

The set of *intralayer edges* is typically defined in the following way. On the *layer* with *co-occurrence perspective element*, *intralayer edges* connect neighbouring sentences, neighbouring words in a sentence and neighbouring syllables in a word. Similarly, for the *shuffle layers intralayer edges* are defined for linguistic units that are neighbouring in the randomized text. On the *layers* with the *syntax perspective element*, *edges* connect neighbouring sentences, neighbouring syllables in words and on the word level *edges* are defined betweeen words adjacent in the syntax dependency tree of each sentence. For an example of the syntax dependency tree on a sentence in Croatian see Figure 15.3 .

Co-occurrence as the network construction principle is sometimes extended to cover neighbouring for a wider window, for more details see e.g. [18]. There are many other candidates for *construction perspect* e.g. *clique*, where one would connect all words in a sentence as a clique, and likewise for other language levels . In another construction one could connect words that differ only in the last syllable [31], etc. Through *language perspect* variuos languages could be compared, but also dialects of the same language as well as development and changes of a language over time.

Interlayer edges betwen *layers* that differ in the *hierarchy perspect* are typically defined on containment bases, connecting a word with syllables it contains, connecting a syllable with graphemes that it contains and so on, as in the *MLN* depicted in Figure 15.1 . Similarly, in some cases edges could be defined between linguistic units that are not necessarily on consequent levels, connecting for example a word with all of the graphemes it contains.

Interlayer edges are defined to reflect the analysis and comparisons of different subsystems of an *MNL* model, reflecting different perspectives one takes when considering one or several languages.

15.4 Diagnostics in *MLN* Model

So far diagnostic for language networks has been reported for isolated subsystems. MLN model has the potential to extent from isolated to integral diagnostics, enabling better insights in mutual interactions of language subsystems. Here we discuss initial steps in this direction.

Various graph and network diagnostics could be applied both to individual *layers* as well as to *underlaying graph* of an *MLN* . For the linguistic analysis one can compare networks and perform relevant diagnostics for the chosen *layers*.

For example, Croatian is generally considered as a mostly free word-order language. In order to support this classification of Croatian language, some conclusions on the importance of word

order in Croatian could be drawn e.g. from comparing layers

$$(croatian, word, syntax) \qquad \text{and} \qquad (croatian, word, shuffle)$$

of the *MLN* model shown in Figure 15.1 .

These listed *layers* differ only in one *perspect*, that is *syntax* v.s. *shuffle*. Comparing *layers* that differ in more than one *perspect* can also be of interest. Since word-order is more strict in English than in Croatian,

one could, for example, compare Croatian text with its English translation, but randomized on the sentence level. An *MLN* model suitable for such comparison could for example be obtained by considering the *MLN* given in Figure 15.1 but with additional *language perspect* $\{croatian, english\}$. One can visualize this model as two coppies of the model from Figure 15.1 , one in Croatian, other in English, with additional interlayer edges as needed. Such *MLN* model would have 18 layers and, in particular, *layers*

$$(croatian, word, syntax) \qquad \text{and} \qquad (english, word, shuffle) \ .$$

would be of interest for the analysis described above.

Kivela et al. [168] review attempts to generalize single-layer-network diagnostics to multilayer networks. This includes e.g. methods of multiway-data-analysis and tensor-decomposition. The later method is based on representing a multilayer network as adjacency tensor .

For the *MNL* model $M = (V_M, E_M, V, L, C)$ adjacency tensor

$$\mathscr{A} \in \{0,1\}^{|V|\times|V|\times|L_0|\times|L_0|\times|L_1|\times|L_1|\times|L_2|\times|L_2|}$$

is such that its tensor element has value 1 if and only if there is the corresponding edge in M, and has value 0 otherwise.

Tensor representation enables one to directly apply methods from the tensor-analysis literature to multilayer networks .

Kivela et al. [168] show how the rank of such a tensor can be reduced. This process of tensor "flattening" leads to the so-called supra-adjacency matrices enabling one to apply known tools and methodology that is used for matrices.

15.4.1 The Network Structure Analysis

We now review some of the most important network measures [67] that can be applied as network diagnostic to an individual *layer* or to the *underlaying graph* of an *MLN*. In this case the individual *layer* or *underlaying graph* of an *MLN* are considered simply as networks or graphs.

A network or graph $G = (V, E)$ is a pair of a set of nodes V and a set of edges E, where N is the number of nodes and K is the number of edges. A network is directed if the edges have a direction associated with them. A network is weighted if there is a weight function ω that assigns value (real number) to each edge.

The density D of a network is defined as a ratio of the number of edges in the network to the number of possible edges. For the directed networks it is calculated using following equation:

$$D = \frac{K}{N(N-1)} \ . \tag{15.2}$$

A path in a network is a sequence of edges which connect a sequence of nodes that are all distinct from one another. A shortest path between two nodes i and j is a path with the shortest length and it is called distance between i and j and is denoted as d_{ij}. The average path length of a

network is defined as the average number of steps along the shortest paths for all possible pairs of network nodes. For the directed network the average path length L is calculated as:

$$L = \sum_{i,j} \frac{d_{ij}}{N(N-1)} \, . \tag{15.3}$$

The clustering coefficient of a node measures the density of edges among the immediate neighbors of a node. For weighted networks the clustering coefficient of a node i is denoted by c_i and defined as the geometric average of the subgraph edges weights:

$$c_i = \frac{1}{k_i(k_i-1)} \sum_{i,j} (\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3} \, , \tag{15.4}$$

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

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

$$C = \frac{1}{N} \sum_{i} c_i \, . \tag{15.5}$$

Transitivity of a network, T, is the fraction of all possible triangles present in the network. Possible triangles are identified by the number of triads (two edges with a shared node):

$$T = \frac{3\#triangles}{\#triads} \, . \tag{15.6}$$

The number of network components is denoted by ω. If $\omega > 1$, C is computed for the largest network component.

Reciprocity of a network, denoted by ρ, is defined as:

$$\rho = \frac{\sum_{i \neq j}(a_{ij} - \overline{a})(a_{ji} - \overline{a})}{\sum_{i \neq j}(a_{ij} - \overline{a})^2} \tag{15.7}$$

where $a_{ij} = 1$ if a edge from node i to j is there, and $a_{ij} = 0$ if not, and the average value $\overline{a} = \frac{\sum_{i \neq j} a_{ij}}{N(N-1)}$.

Above measures are used for the results of the experiments we conducted as is presented in the next Section in Tables 15.1 and 15.2.

15.4.2 Experiments and Results

Multilayer Croatian (HR) and English (EN) networks are constructed from five variations of the same collection of texts (HOBS and PENN corpora): three on the word-level (syntax, co-occurrence and its shuffled counterpart) and two on the sub-word level (syllables and graphemes). More clearly, five different realizations of the very same text in one language are used to construct network layers (all weighted and directed) using five different relationships among linguistic units: syntax (SIN), co-occurrence (CO), shuffled (SHU), syllables (SYL) and graphemes (GR).

The five variations of the text described above, naturally fit the *MLN* model, similar to the model given in Figure 15.1 , but with English language *perspective element* added. Such *MLN* model has 18 layers, but in this experiment is reduced to 10 non-empty layers, since for the syllabic and graphemic *layer* we consider only the co-occurrence construction principle.

The data for multilayer Croatian network is derived from Croatian Dependency Treebank [257]. The corpus size is currently 3,465 sentences (88,045 tokens). Dataset from the Penn Treebank corpus [258] is used for the English multilayer network . The dataset contains 3,829 sentences

HR	CO	SIN	SHU	SYL	GR
N	23359	23359	23359	2634	34
K	71860	70155	86214	18849	491
D	0.00026	0.00026	0.00032	0.0054	0.875
L	4.01	1.81	3.74	1.86	1.58
C	0.167	0.120	0.182	0.255	0.636
T	0.004	0.003	0.013	0.120	0.522
ω	2	2	2	17	1
ρ	0.049	0.041	0.085	0.139	0.531

Table 15.1: Measures (N no. of nodes, K no. of edges, D density, L avg. path length, C clust. coeff., T transitivity, ω no. of components, ρ reciprocity) for co-occurrence (CO), syntax (SIN), shuffled (SHU), syllable (SYL) and grapheme (GR) network layers in Croatian.

(94,084 tokens). Syntax Dependency Tree is a tree parsed from original sentence according to the syntax relationships among words . For this work we use the text of original sentences for the construction of co-occurrence [18] and shuffled layers [21], and syntax relationships from treebank corpora for the syntax layer. Further, we decompose Croatian and English words to syllables and graphemes.

EN	CO	SIN	SHU	SYL	GR
N	10930	10930	10930	2599	26
K	50299	52221	58920	6053	333
D	0.00084	0.00087	0.00099	0.0018	1.025
L	3.465	1.959	0.454	1.876	1.511
C	0.286	0.153	0.295	0.057	0.838
T	0.009	0.014	0.016	0.020	0.654
ω	3	3	1	54	1
ρ	0.051	0.046	-0.0005	0.017	0.575

Table 15.2: Measures (N no. of nodes, K no. of links, D density, L avg. path length, C clust. coeff., T transitivity, ω no. of components, ρ reciprocity) for co-occurrence (CO), syntax (SIN), shuffled (SHU), syllable (SYL) and grapheme (GR) network layers in English.

The standard network measures of all five network layers are given in Table 15.1 for Croatian and Table 15.2 for English.

The number of nodes (N) on the word-level layers is preserved in both languages (HR: 23359, EN: 10930). On subword-levels the inventory of linguistic units is smaller (around 2500 syllables and 30 graphemes per language) which disables direct comparisons of network measures at word and subword-level. Still some additional remarks are worth noticing. The average path length (L) decrease from co-occurrence to syntax, as expected, but interestingly it is of the same range for the syntax and syllabic layer of both languages . The clustering coefficient (C) (obtained from the undirected versions of the same networks) increases on the syllabic sub-word level for Croatian and decreases for English. At the same time English layers are more clustered than the Croatian ones.

One of the explanations for this difference can be found in the number of connected components (ω) which is three times higher for English then for Croatian syllabic layer, see Tables 15.1 and

15.2. Grapheme layers of both languages expectedly, exhibit the highest clustering coefficients. Moreover, the Croatian syllabic layer has higher reciprocity then the corresponding word layers for one order of magnitude. The graphemic layers of both languages exhibit peculiar features due to the small number of nodes or in other words due to the high density (0.88 - HR; 1.03 - EN).

The transitivity (T) shows constant increase across layers (from CO to GR) regardless of the language. The same holds for density (D). Additionally, transitivity and density of English layers are consistently higher then the corresponding Croatian values. This is caused by the high flectivity of Croatian language, which results in a bigger inventory of linguistic units.

15.5 Conclusions and Future Work

Multilayer networks and related concepts have been used for the description and analysis of various complex systems in many fields, such as for example biological, physical, social and information sciences, for an overview see [168]. These are the first steps in the work on a multilayer network model for languages.

MLN model is universal enough to allow extensions with additional perspects as needed. Indeed, in the future we plan to extend the model with as many perspects as is linguistically required to quantitatively study the structure of language in a unified framework of *MLN*.

From the point of view of diagnostics, *MLN* model allows various approaches. The most obvious approach is through the field of graph theory and network analysis by comparison of different *layers*. Similar analysis can be applied to the *underlying graph* of an *MLN*. More characteristics of languages could be obtained through tensor analysis.

Proposed *MLN* model can be of high relevance for computer science as well, especially for applications which process natural language or retrieve information. For instance in text summarization, text quality assessment, keyword extraction etc.

All of the above approaches and results could possibly be combined to extract conclusions about syntax, semantics and overall complexity of language. We intend to pursue research in this direction in the future.

16. Multilayer Network of Language: a Unified Framework for Structural Analysis of Linguistic Subsystems

16.1 Abstract

Recently, the focus of complex networks' research has shifted from the analysis of isolated properties of a system toward a more realistic modeling of multiple phenomena - multilayer networks. Motivated by the prosperity of multilayer approach in social, transport or trade systems, we propose the introduction of multilayer networks for language. The multilayer network of language is a unified framework for modeling linguistic subsystems and their structural properties enabling the exploration of their mutual interactions. Various aspects of natural language systems can be represented as complex networks, whose vertices depict linguistic units, while links model their relations. The multilayer network of language is defined by three aspects: the network construction principle, the linguistic subsystem and the language of interest. More precisely, we construct a word-level (syntax, co-occurrence and its shuffled counterpart) and a subword-level (syllables and graphemes) network layers, from five variations of original text (in the modeled language). The analysis and comparison of layers at the word and subword-levels is employed in order to determine the mechanism of the structural influences between linguistic units and subsystems. The obtained results suggest that there are substantial differences between the networks' structures of different language subsystems, which are hidden during the exploration of an isolated layer. The word-level layers share structural properties regardless of the language (e.g. Croatian or English), while the syllabic subword-level expresses more language dependent structural properties. The preserved weighted overlap quantifies the similarity of word-level layers in weighted and directed networks. Moreover, the analysis of motifs reveals a close topological structure of the syntactic and syllabic layers for both languages. The findings corroborate that the multilayer network framework is a powerful, consistent and systematic approach to model several linguistic subsystems simultaneously and hence to provide a more unified view on language.

16.2 Introduction

Recently, the field of complex networks has shifted from the analysis of isolated network (capturing and modeling one aspect of the examined system) toward the analysis of the family of complex

networks simultaneously modeling different phenomena (aspects) of the examined system, or modeling interactions and relationships among different subsystems. The rise of this more realistic framework for a complex network analysis considers different layers, levels or hierarchies for different aspects of the system. In other words, multiple phenomena are characterized by multiple types of links across various levels of representations or various dimensions of relations for multiple subsystems. The multilayer network approach has been addressed in the analysis of real international trade analysis [253], social interactions in the massive online game [254], web-search queries [247], in transport and infrastructure [246,248,252,260] and in the examination of the brain's function [255]. There are variations in formal representation of the multilayer networks [246, 261], multidimensional networks [247], multiplex networks [248,249,252], interdependent networks [250, 260] and networks of networks [251, 259]. A thorough discussion that compares, contrasts, and translates between notions of multilayer, multiplex, interdependent networks and networks of networks is in [168], which together with [169] presents an detailed overview of multilayer network theory.

Viewed as a unique, biologically-based human faculty [262], language has been recognized as the reflection of the human cognitive capacities, both in terms of its structure and its computational characteristics [263]. Studying languages at intra- and cross-linguistic levels is of paramount importance in relation to our biological, cultural, historical and social beings. Hence, human languages, besides still being our main tools of communication, reflect our history and culture. Language can be seen as a complex adaptive system [264], evolving in parallel with our society [265] .

Various aspects of natural language systems can be represented as complex networks, whose vertices depict linguistic units, while links model their morphosyntactic, semantic, pragmatic, etc. interactions. Thus the language network can be constructed at various linguistic levels: syntactic, semantic, phonetic, syllabic, etc. So far there have been efforts to model the phenomena of various language subsystems and examine their unique function through complex networks . Still, the present endeavors in linguistic network research focus on isolated linguistic subsystems lacking to explain (or even explore) the mechanism of their mutual interaction, interplay or inheritance. Obtaining such findings is critical for deepening our understanding of conceptual universalities in natural languages, especially to shed light on the cognitive representation of the language in the human brain [266].

Therefore, one of the main open questions in linguistic networks is explaining how different language subsystems mutually interact [264, 267]. The complexity of any natural language is contained in the interplay among several language levels. Below the word-level, it is possible to explore the type of phonology, morphology and syllabic subsystem complexity . For example, the phonology subsystem complexity is reflected in the morphology subsystem complexity. On the word-level, the morphology subsystem complexity reflects in the complexity of the word order, syntactic rules and the ambiguity of lexis . Since the word order can be considered as the primary factor (but not the only one) that determines linguistic structure, it is important to explore the subsystems' interactions by which it is influenced.

In this research we use the multilayer network framework to explore the structural properties of various language subsystems and their mutual interactions. The multilayer network of a language is constructed for the word (co-occurrence, syntax and shuffled) and subword (syllables and graphemes) language levels . The systematic exploration of layers properties is presented for the Indo-European family of languages: one representative of the Slavic group - Croatian, and one representative of the Germanic group - English. The analysis and comparison of layers is employed in order to determine structural influences and trade-offs between the subsystems of language.

Our work contributes mainly to the field of linguistic network research by proposing the multilayer network model for language. The multilayer language network model is established on three aspects: the network construction principle, the linguistic subsystem and the language

of interest. Moreover, we introduce the preserved weighted overlap as the measure of word-level layers similarity in weighted and directed networks. Finally, we propose the characterization of word vs. subword layers relationships by correlations of triad significance profiles, as a possible quantification of the inter layer relationships.

16.2.1 Related Work

The Language Networks

The pioneering work of Dorogovtsev and Mendes [7] describes language as a self–organizing network of linked words. The observed word web structure distributions naturally emerge from the evolutionary dynamics. Masucci and Rodgers [11] investigate the topology of Orwell's 1984 within the framework of complex network theory. They exhibit local preferential attachment as growth mechanisms of written language and the allocation of a set of preselected vertices that have a structural rather than a functional purpose. Choudhury and Mukherjee in [6] provide a suitable framework to model a language from three different perspectives microscopic (utterances), macroscopic (grammar rules and a vocabulary) and mesoscopic (linguistic entities - letters, words or phrases). The authors mainly present an overview of the structure and dynamics at the mesoscopic level. Sole et al. [22] review the state-of-the-art on language networks and their potential relevance to cognitive science. They also consider the intertwining of language levels related to multiple layers of complexity in terms of the networks of connected words in order to shed light onto the relevant questions concerning language organization and its evolution. In [210] Cong and Liu provide an extensive insight into the language networks which positions human language as a multi-level system in the discipline of complex network analysis. Relationships between the system-level complexity of human language (determined by the topology of linguistic networks) and microscopic linguistic features (as the traditional concern of linguistics) are positioned within a holistic quantitative approach for linguistic inquiry, which contributes to the understanding of human language at different granularities.

The Word-level Networks: Co-occurrence vs. Shuffled

The construction of language networks relies on the well-established principles of modeling word interaction from the word order in a sentence or in short from their co-occurrence in text . A substantial part of reported research on language networks is dedicated to a detailed structural analysis of co-occurrence networks interpreting their topological properties in the linguistic context [6, 7, 10, 11, 18]. Thus, in the linguistic co-occurrence networks properties are derived directly from the word order in texts by connecting words within a window of certain size or sentence. Still, 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 [21] and networks constructed from syntax dependencies in texts.

The Word-level Networks: Syntax

The syntactic structure of language is captured through syntax dependency relations between a pair of words in a sentence: the head word -– the governor of relationship and the dependent word - the modifier . Syntax dependencies between words are formally expressed by dependency grammar (e.g. a set of productions (rules) in the form of a grammar). The dependency grammar is used to parse the syntactic relationships from a sentence in the form of a syntax dependency tree . Thus, the syntax dependency treebank is the set of syntax dependency trees parsed from the sentences in a corpus . Ferrer i Cancho et. al [23], in the seed work on syntax complex networks model the syntactic dependency relationships of three languages comparatively (Czech, German, Romanian). The set of analyzed languages is extended to 7 in [61], comparing the structure of global syntactic dependency networks. The results in [60, 61] show that the proportion of syntactically incorrect

relationships rises from about 30 % to a high 50 % in a co-occurrence networks constructed with a window of size 2 and 3 respectively. In [17], based on the comparison of one syntactic dependency network and two co-occurrence networks of Chinese, the authors confirm small-world and scale-free properties, suggesting that scale-free architecture is of essential importance to the syntax subsystem of human language. Liu et al. [24] and Abramov and Meheler [56] use network parameters derived from the syntax relationships for hierarchical clustering of languages, deriving the model of the genealogical similarity among 15 and 11 languages respectively. The obtained results on syntax networks suggested that a natural approach to modeling human language is considering the structure of the syntactic dependency relationships besides the simple word-order relations reflected in co-occurrence networks. Amancio et al. [268] explore the Portuguese syntax dependencies for automatic summarization of the news.

The Subword-level Networks: Syllables

The coherent results from language networks involving units smaller than words , such as syllables [30, 44], phonemes [26] or morphemes [56] are still missing. Morphological networks for English and German are presented in [56] and the network properties are expressed in terms of graph entropy measures. So far, syllable networks have been constructed exclusively for Portuguese [30] and Chinese [44] . Syllables are a natural intermediate level in the analysis of spoken (as opposed to written) language, since they carry prosody during pronunciation. The investigation of syllables is particularly interesting for their role in language acquisition. Children begin to learn language through syllables, culminating in the development of their mental lexicons [266, 269]. The model of language acquisition was recreated with humanoid robots using syllables as basic units [270] or by artificial agents [271]. Both studies witness the complexity of a language syllable system as an important factor in language acquisition.

The Subword-level Networks: Graphemes

Language is written with a set of abstract orthographic symbols (letters of an alphabet) – graphemes. Graphemes are the smallest semantically distinguishing units (the basic linguistic units) in a written language, analogous to the phonemes in spoken language . The complex networks of grapheme subsystem of language have been studied sporadically [272]. Kello and Beltz analyzed the structure of the complex network constructed from the orthographic wordform lexicon, where words are connected if one is a substring of the other. Phonemes have attracted more attention since many psycholinguistics studies regarding the representation of mental lexicon used for speech production, word recognition and language processing have been reported [25, 273–276]. Phonetic networks are typically constructed from words in a lexicon, establishing links among phonetically similar words – differing in one phoneme.

Network Motifs for Language

Motifs are subgraphs defined as simple building blocks of directed complex networks [51] . Motifs are used to detect the structural similarities and differences between networks on the local level. In [52] the significance profiles of motifs derive several superfamilies of networks - the language networks forms one supra family based on the triad significance profile. Binemann et al. in [2] use motifs to quantify the differences between natural and generated language. The frequencies of three-vertex and four-vertex motifs for six languages are compared with the generated language from n-gram statistical model (n-grams are a sequence of n units from a given text). The authors show that the four-vertex motifs are directly interpretable by semantic relations of polysemy and synonymy. An initial attempt to analyze undirected triads in a multiplex network, by representing positive and negative social interactions of game players in massive online game is reported in [254].

The linguistic features of Croatian and English

A short recapitulation of the main properties of the Croatian and English languages establishes the linguistic framework needed for the comparison across languages as well as for the interpretation of insights into their structural characteristics. Croatian is a highly flective Slavic language and words can have seven different cases for singular and seven for plural, genders and numbers. The Croatian word order is mostly free, especially in non-formal writing. These features place Croatian among morphologically rich and mostly free word-order languages. English grammar has minimal inflection compared with most other Indo-European languages, therefore it is considered to be analytic. English word order is almost exclusively subject-verb-object. Both languages are characterized by an accentuation system developed on syllables.

English has been studied extensively in a complex networks framework [2, 6, 10, 11, 22, 24], still no systematic effort explaining the effects of mutual interaction of different subsystems has been reported. So far the Croatian has been quantified in a complex networks framework based on the word co-occurrences [10, 18] and compared with shuffled counterparts [21]. The syntax relationships of Croatian as well as syllabic subword units are novelty characterized through the lenses the analysis of complex networks in this research.

16.3 Methods

More details about complex networks analysis and the definition of measures can be found in [67]. Here we list a short definition of measures needed for the exploration of network layers. The network $G = (V,E)$ is a pair of a set of vertices V and a set of links E, where N is the number of vertices and K is the number of links. In weighted networks every link connecting two vertices i and j has an associated weight w_{ij}. The number of network components is denoted by ω .

For every two connected vertices i and j the number of links lying on the shortest path between them is denoted as d_{ij}, then the average path length between every two vertices i, j is $L = \sum_{i,j} \frac{d_{ij}}{N(N-1)}$. If the number of components $\omega > 1$, L is computed for the largest connected component in network. If in a directed network there is no path between two vertices, then shortest path between two vertices is assigned to 0.

For weighted networks the clustering coefficient of a vertex i is defined as the geometric average of the subgraph link weights: $c_i = \frac{1}{k_i(k_i-1)} \sum_{j,k} (\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3}$, where k_i is the degree of the vertex 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 coefficient of a network is defined as the average value of the clustering coefficients of all vertices in an undirected network: $C = \frac{1}{N}\sum_i c_i$.

The transitivity of a network is the fraction of all possible triangles present in the network. Possible triangles are identified by the number of triads (two links with a shared vertex): $T = (3\#triangles)/(\#triads)$.

The in-degree and out-degree $k_i^{in/out}$ of vertex i is defined as the number of its in and out nearest neighbors. The in-strength and the out-strength $s_i^{in/out}$ of the vertex i is defined as the number of its incoming and outgoing links, that is: $s_i^{in/out} = \sum_j w_{ji/ij}$.

The in- and out- selectivity of the vertex i is then defined as proposed in [11]:

$$e_i^{in/out} = \frac{s_i^{in/out}}{k_i^{in/out}}. \tag{16.1}$$

The power-law distribution is defined as: $P(k) \sim k^{-\gamma}$ where γ is the power-law exponent .

16.3.1 Network Motifs Analysis

Network motifs are connected and directed subgraphs (of three to up to eight vertices) occurring in complex networks at numbers that are significantly higher than those in randomized networks with

the same degree distribution [51,52]. Here, we analyze only triads (all possible directed three-vertex subgraphs) by calculating their frequencies, Z-scores and triad significance profiles (TSP).

The scores Z_i for each triad i is calculated using equation:

$$Z_i = \frac{N_i^{orig} - \langle N_i^{rand} \rangle}{\sigma_i^{rand}}, \tag{16.2}$$

where N_i^{orig} is the count of appearances of the triad i in the original network, while $\langle N_i^{rand} \rangle$ and σ_i^{rand} are the average and the standard deviation of the counts of the triad i over a sample of randomly generated networks.

The triad significance profile TSP is the normalized vector of statistical significance scores Z_i for each triad i $TSP_i = \frac{Z_i}{\sqrt{\sum_i Z_i^2}}$.

16.3.2 The Multilayer Network

Since language networks can be viewed through different aspects: different levels (e.g. word-level, subword-level), different construction rules (e.g. co-occurrence, shuffle), different languages, etc. there is a need for a general network model that can capture all these aspects in one single framework. Therefore, we propose an application of general multilayer networks model introduced by Kivelä *et al.* in [168] to the multilayer language networks.

According to [168], a multilayer network can have any number d of aspects defined as a sequence $L = \{L_a\}_{a=1}^{d}$. There is one set of elementary layers L_a for each aspect a. In a multilayer network it is possible to construct a set of layers by assembling a set of all of the combinations of elementary layers using a Cartesian product $L_1 \times ... \times L_d$.

The multilayer network is a quadruplet $M = (V_M, E_M, V, L)$, where $V_M \subseteq V \times L_1 \times ... \times L_d$ that contains only the vertex-layer combinations in which a vertex is present in the corresponding layer, and where E_M is a set of pairs of the possible combinations of vertices and elementary layers, $E_M \subseteq V_M \times V_M$. V is a set of all vertices in all layers. Multiplex is a special case of multilayer network, which satisfies the condition that the set of vertices is shared across layers. Thus, in a multilplex network inter layer connections between different layers have 1:1 or 0:1 cardinality of relationships.

Next we present equations for the calculation of the overlap between two layers. These equations can be applied only to a multiplex network, when two layers share the same vertices (e.g. in our case it is applicable only to the construction aspects of the word-level layers in one language). In the following text we use only α and α' for the shorter notation of the one layer in the multilayer network.

Jaccard index for link overlap between two network layers α and α' is :

$$J(E_\alpha, E_{\alpha'}) = \frac{|\, E_\alpha \cap E_{\alpha'} \,|}{|\, E_\alpha \cup E_{\alpha'} \,|}. \tag{16.3}$$

In the same way we can calculate the Jaccard index for weight overlap (W).

The preserved weighted ratio on intersected links between network layers α and α' is (modified from total weighted overlap [249]) is :

$$PW(E_\alpha, E_{\alpha'}) = \sum_{i,j} \frac{\min(w_{ij^\alpha}, w_{ij^{\alpha'}})}{\max(w_{ij^\alpha}, w_{ij^{\alpha'}})}. \tag{16.4}$$

The preserved weighted overlap (WO) is a normalized preserved weighted ratio :

$$WO(E_\alpha, E_{\alpha'}) = \frac{PW(E_\alpha, E_{\alpha'})}{|\, E_\alpha \cap E_{\alpha'} \,|}. \tag{16.5}$$

16.3.3 Croatian and English Datasets

The data sets for multilayer Croatian networks are derived from the HOBS corpus - the first version of the Croatian Dependency Treebank [257]. HOBS is extracted as a part of the Croatian National Corpus [258] and annotated at the analytical layer following the Prague Dependency Treebank formalism adapted to Croatian. The corpus size is currently 3,465 sentences (88,045 tokens) .

The English dataset contains 3,829 sentences (94,084 tokens) from the Penn Treebank corpus [277, 278]. The size of the extracted Penn subset is intentionally of the same size as HOBS in order to allow for systematic comparisons across the layers, constructed from comparable corpora of different languages .

Multilayer Croatian (HR) and English (EN) networks are constructed from five variations of HOBS and Penn corpora: three on the word-level (syntax, co-occurrence and it's shuffled counterpart) and two on the subword-level (syllables and graphemes). More clearly, five different realizations of the very same text in one language are used to construct the network layers (all weighted and directed) using five different relationships among the linguistic units: syntax (SIN), co-occurrence (CO), shuffled (SHU), syllables (SYL) and graphemes (GR) .

16.3.4 Language Networks Construction

The language networks construction principle arises form the vary nature of text (and speech), which is always advancing in an onward direction, hence to use directed and weighted links representing relations among linguistic units [11, 18, 210]. The co-occurrence relation is established between two adjacent words within a sentence (CO), where the direction of link reflects the words sequencing and weight on the link reflects the frequency of words-pair mutual appearance .

The syntax relationships among word-pairs are parsed from the HOBS and Penn, as well as the text of the original sentences [277]. The sentences' boundaries are preserved, since the syntax dependency is inherent to the sentence (SIN). Thus, the sentence boundaries are considered as linkage delimiters for the co-occurrence layers as well .

Next, the original text is shuffled in order to obtain a shuffled counterpart (SHU), again considering the sentence boundaries. Commonly, the shuffling procedure randomizes the words in the text, transforming the text into a meaningless form. We shuffled the words within the original sentences, preserving the vocabulary size, the word and sentence frequency distributions, the sentence length (the number of words per sentence) and sentence order [21]. Figure 16.1 (top part) presents the principles of word-level layers' construction for one sentence .

Next, we use the Croatian syllabification with a maximal onset algorithm to prepare the last data set – syllables [280], again from the words in the original sentences. The English syllables are obtained from the dictionary with syllabified words [279]. The process omitted words which were not contained in the syllabified dictionary. The syllable layers are constructed from the co-occurrence of syllables within words (SYL) - presented at the (D) part of Figure 16.1 .

Finally, we consider the set of graphemes present in words, where graphemes (GR) represent the most elementary subsystem of each language - orthographical. Since, there are some foreign words present in the used corpora we preserved the original orthographic symbols, resulting in a slightly larger number of graphemes (e.g. in Croatian foreign names contain original diacritic symbols, so we obtained q, w, x, y as Croatian graphemes as well) .

Multilayer language network for this work can be defined with the set L of three aspects: construction L_1, linguistic subsystem L_2 and language L_3, where L_1 ={*co-occurrence, syntax, shuffle*}, $L_2 = \{word, syllable, grapheme\}$ and $L_3 = \{Croatian, English\}$. Therefore, it is possible to have 18 different layers in total ($3x3x2$), although not all the layers are of equal interest. More precisely, one can note that some layers are equal due to the specific construction rules. Since we connect only neighboring syllables within the word, all three layers *(co-occurrence, syllable, Croatian)*, *(syntax, syllable, Croatian)* and *(shuffle, syllable, Croatian)* are equal. The same holds

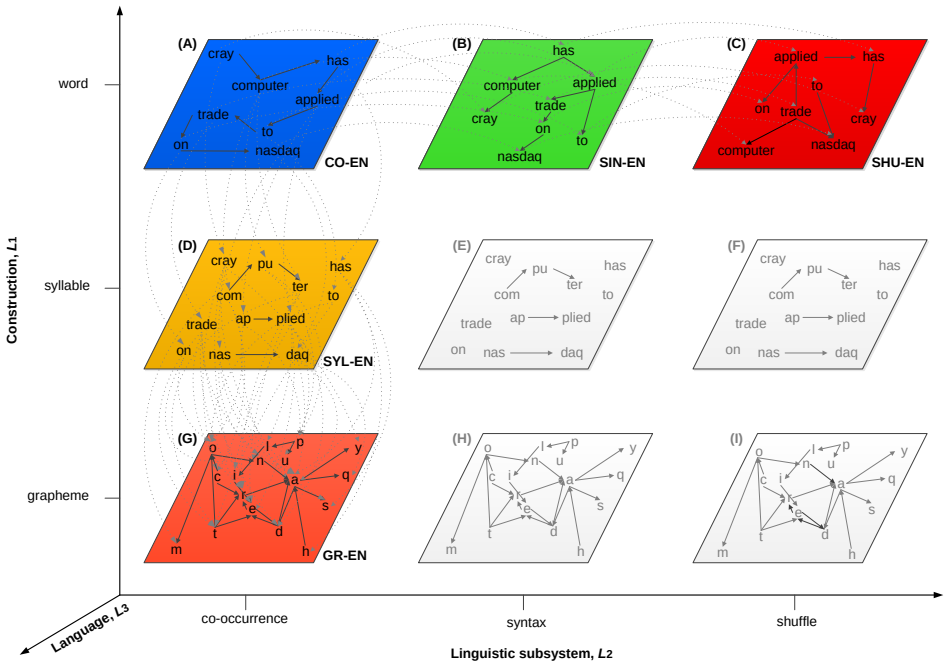


Figure 16.1: **The multilayer language network.** Three word-level layers: (A) co-occurrence; (B) syntax; (C) shuffled; and two subword-level layers: syllables (D) and graphemes (G) constructed from the English sentence "*Cray Computer has applied to trade on NASDAQ.*"; according to three aspects of multilayer network model of language: construction, linguistic subsystem and language. Note- layers (E) and (F); (H) and (I) are gray, since they are disregarded in analysis (identical with layers (D) and (G) respectively).

for English syllables, and for graphemes in both languages as well, as shown gray for (E), (F), (H) and (I) parts of Figure 16.1.

It is worth noticing, that the word-level layers are forming the multiplex networks (have 1:1 inter- connections), while the connections between word and subword layers are not coupled (have N:M inter- connections).

To sum up, in total we construct ten layers: five of Croatian *(syntax, word, Croatian), (co-occurrence, word, Croatian), (shuffle, word, Croatian), (co-occurrence, syllable, Croatian), (co-occurrence, grapheme, Croatian)* and five of English language *(syntax, word, English), (co-occurrence, word, English), (shuffle, word, English), (co-occurrence, syllable, English), (co-occurrence, grapheme, English)*, with shortened notations: SIN-HR, CO-HR, SHU-HR, SYL-HR, GR-HR, SIN-EN, CO-EN, SHU-EN, SYL-EN and GR-EN.

Multilayer 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 [8]. The frequencies and triad significant profiles of motifs are obtained with the FANMOD tool [55].

16.4 Results

Initially we explore the characterization of all isolated layers with the standard set of network measures (see Methods Section). The results for all ten network layers (for both languages) are in Table 16.1. The networks are of different sizes, however, the presented measures (L, C and T) are all normalized (by the number of nodes). More important, all networks for one language

are constructed form the same text and that gives us the possibility to compare various linguistic realizations of the same source text. Even the datasets for two languages contain approximately the same number of sentences, allowing for comparisons between languages as well. The average path length (L) decrease from co-occurrence to syntax, as expected, but interestingly it is of the same range for the syntax and syllabic layer. The clustering coefficient (C) (obtained from the undirected versions of the same networks) increases on the syllabic subword-level for Croatian and decreased for English. The clustering of English CO and SHU word-levels are higher than their Croatian counterparts. Still, clustering coefficients of SIN layers in both languages are of the same range.

Also, the Croatian syllabic layer has the transitivity higher than the corresponding word layers by one order of magnitude. The numbers of connected components in SYL layers are the highest compared with other layers, and three times higher for English than Croatian. The graphemic layers of both languages exhibit peculiar features due to the small number of vertices, or in other words, due to the high density of GR networks (0.9 - HR; 1.03 - EN).

	CROATIAN					**ENGLISH**				
	CO	SIN	SHU	SYL	GR	CO	SIN	SHU	SYL	GR
N	23359	23359	23359	2634	34	10930	10930	10930	2599	26
K	71860	70155	86214	18849	491	50299	52221	58920	6053	333
L	4.01	1.81	3.74	1.86	1.58	3.47	1.96	0.45	1.88	1.51
C	0.167	0.120	0.182	0.255	0.636	0.286	0.153	0.295	0.057	0.838
T	0.004	0.003	0.013	0.120	0.522	0.009	0.014	0.016	0.020	0.654
ω	2	2	2	17	1	3	3	1	54	1

Table 16.1: **The standard network measures for ten layers. Measures** (N no. of vertices, K no. of links, L avg. path length, C clust. coeff., T transitivity, ω no. of components) for co-occurrence (CO), syntax (SIN), shuffled (SHU), syllable (SYL) and grapheme (GR) network layers in Croatian and English.

16.4.1 Word-level Layers

For the word-level layers we initially examine the distributions. Figure 16.2 shows the rank distributions for in- and out- degrees of word-level layers in both languages. The exploitation of the same data source per each language caused the high overlap of exposed distributions. Analogously, the in- and out- strength distributions are overlapped as well, for both languages. The power-law γ coefficients for all distributions of word-level layers are in a range between 2.14 and 2.49; thus CO, SIN and SHU layers exhibit the power-law distributions for degree and strength regardless of the language.

The potential of selectivity (in- and out-) to differentiate between different text types [71] or the ability to extract keywords [59] (identifying and ranking the most representative features of the source text) is restated in this work for the differentiation of language layers as well. Figure 16.3 reveals that the rank distributions of in- and out- selectivity for all word-level layers are apart. Selectivity distributions of all three layers co-occurrence (CO), syntax (SIN) and shuffled (SHU) are separated for both languages.

The correlation matrices in Figure 16.4 show the intra (CO-CO, SHU-SHU and SIN-SIN) and inter layer (CO-SHU, CO-SIN and SHU-SIN) correlations in terms of in- and out- degree, in- and out- strength, in- and out- selectivity distributions. The correlation values for syntax layers of both languages are lower than the corresponding values for the co-occurrence and shuffled layers. Notably, the degree and strength correlation values are higher than the selectivity ones, regardless of the language and layer. Furthermore, Croatian is characterized by higher intra and inter layer

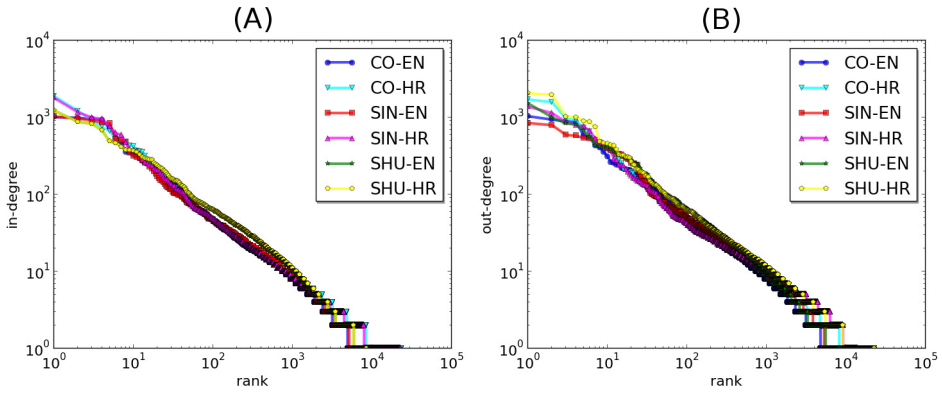


Figure 16.2: **The word-level layers degree rank distributions.** Rank distributions of in- (A) and out- (B) degrees for word-level layers: co-occurrence (CO-HR, CO-EN), syntax (SIN-HR, SIN-EN) and shuffled (SHU-HR, SHU-EN).

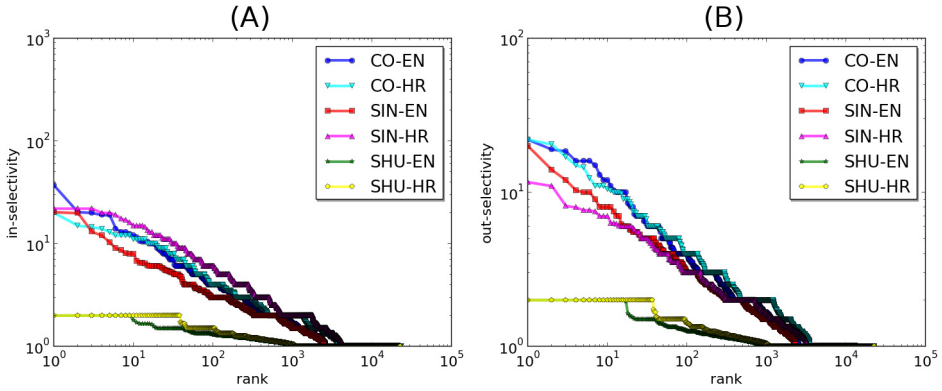


Figure 16.3: **The word-level layers selectivity rank distributions.** Rank distributions of in- (A) and out- (B) selectivity for word-level network layers: co-occurrence (CO-HR, CO-EN), syntax (SIN-HR, SIN-EN), shuffled (SHU-HR, SHU-EN).

correlations than English.

In order to obtain a deeper insight into word-level inter layer relationships we calculated the Jaccard overlap percentage, the percentage of total overlapped weight (W) and the percentage of the preserved weighted overlap (WO) for the overlapping links between word-level layers pairwise (Table 16.2) . The highest percentage of overlapped links is inherent for the intersection of the co-occurrence and syntax layer in both languages, while the overlaps with shuffled layer are expectedly, lower . Furthermore, for both languages the percentage of preserved overlapped weights is relatively high, although slightly lower for English, bearing in mind that less than 20% of the total possible weights on the total intersected links are preserved.

16.4.2 Subword-level vs. Word-level Layers

Subword-level layers syllabic (SYL) and graphemic (GR) in both languages exhibit the power-law γ coefficients between 1.7 and 4.42, which is broader than the observed range of the word-level layers .

If we compare the syllabic layers of both languages, it is possible to notice some differences between Croatian and English. English syllables are characterized by distributions closer to the word-level layers distributions (γ coefficients between 1.87 and 2.14). The Croatian syllabic layer distributions reveal some deviations (γ coefficients are lower - between 1.72 and 1.94). The

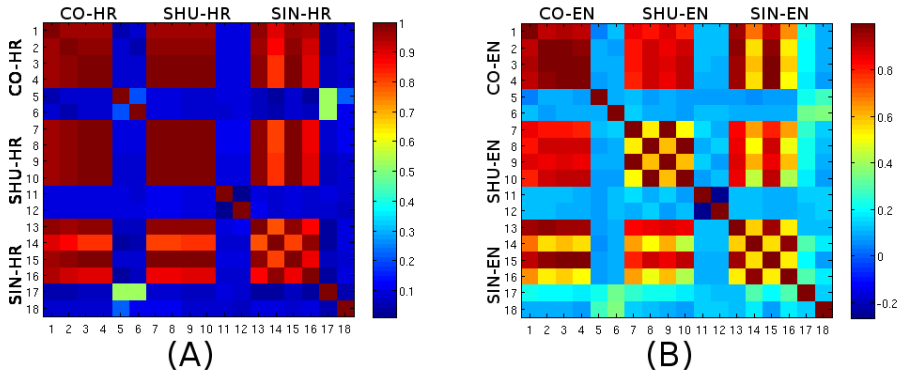


Figure 16.4: **Intra and inter layers correlations matrices.** The correlations matrices for Croatian (A) and English (B): in- out- degree, in- & out- strength and in- & out- selectivity respectively, presenting inter and intra layer correlations for co-occurrence (CO), shuffled (SHU) and syntax (SIN) word-level layers (all p-values ≤ 0.001).

	CROATIAN			**ENGLISH**		
	CO - SIN	CO - SHU	SIN - SHU	CO - SIN	CO - SHU	SIN - SHU
Jaccard	16.72 %	5.47 %	4.81 %	13.44 %	6.31 %	5.34 %
W	18.96 %	6.43 %	5.63 %	13.58 %	6.28 %	4.82 %
WO	90.6 %	76.6 %	74.6 %	90.00 %	74.72 %	73.81 %

Table 16.2: **Overlap of word-level layers.**The Jaccard overlap percentage, total weighted overlap percentage (W) and preserved weighted overlap percentage (WO) between word-level layers (pairwise) for Croatian and English.

grapheme layers have γ coefficients between 1.7 and 4.16 for Croatian and 2.34 and 4.11 for English.

However, in the multilayer language networks it is interesting to take additional insights of the inter layer relationships, mainly to explore the relationships between word vs. subword layers. For this purpose we introduce the analysis of motifs. We exploited the motif frequencies as well as the normalized triad significance profiles (TSP) of all layers for the analysis. The Pearson correlations for all pairs of network layers in Table 16.3 highlight that motif's frequencies in all layers, with the exception of the graphemic layer are correlated. Correlations of normalized TSP indicate that SIN and SYL layers in both languages and additionally for English also CO and SIN layers expose similarities. In order to obtain a deeper insight the normalized significance profiles for CO - SIN - SYL layers of Croatian and English per 13 triadic motifs are compared in Figure 16.5.

16.5 Discussion

The presented findings show that standard network measures on isolated layers exhibit no substantial differences across layers, only slight variations between word and subword-levels. Although, if we compare the structural differences across the examined languages there are indications of different principles in their organization. For instance, English is characterized by higher clustering, with the exception of the syllabic layer. The English syllabic layer has 54 components, while Croatian has 17, which is reflected in the low clustering coefficient of English syllables. This is caused by high flectivity of Croatian, where many words share the suffix - the last syllable, which decreases the number of components, and increases the clustering coefficient. This observation raises a question,

	CROATIAN		ENGLISH	
	Freq.	TSP	Freq	TSP
CO-SHU	**0.99**	0.01	**0.93**	-0.26
CO-SIN	**0.95**	0.42	**0.91**	**0.92**
CO-SYL	**0.96**	-0.03	**0.86**	0.73
CO-GR	-0.18	-0.26	-0.30	0.39
SHU-SIN	**0.93**	0.39	**0.84**	0.04
SHU-SYL	**0.96**	0.32	0.74	0.35
SHU-GR	-0.15	-0.28	-0.17	-0.27
SIN-SYL	**0.95**	**0.83**	**0.99**	**0.91**
SIN-GR	-0.20	-0.21	-0.31	0.28
SYL-GR	-0.21	-0.15	-0.33	0.12

Table 16.3: **The Pearson correlations of triad frequencies and normalized triad significance profiles.**The Pearson correlations of triad frequencies and normalized triad significance profiles (TSP) for all pairs of network layers (co-occurrence (CO), syntax (SIN), syllables (SYL) and graphemes (GR) for Croatian and English (all p-values ≤ 0.001, emphasized values ≥ 0.8).

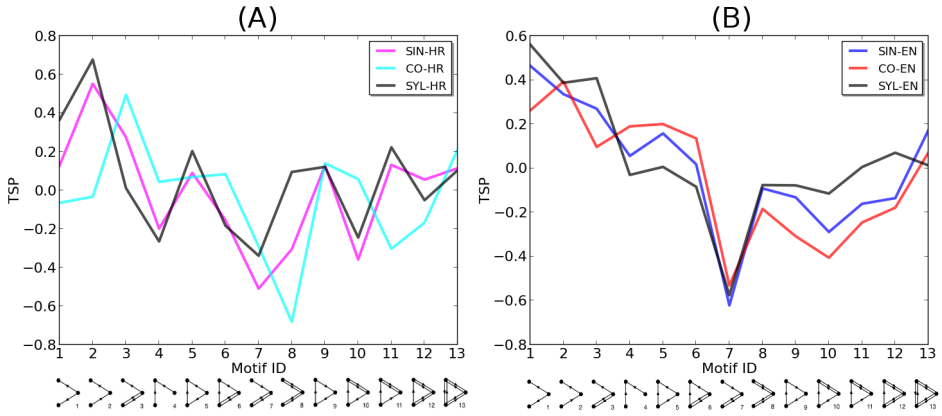


Figure 16.5: **The normalized triad significance profiles.** The normalized triad significance profiles for 13 triadic motifs following the enumeration in [52] comparing co-occurrence (CO), syntax (SIN) and syllables (SYL)layers for Croatian (A) and English (B).

which properties will the morpheme language subsystem expose during the incorporation into a multilayer language framework?

Even a standard distribution analysis is not sufficient to take a deeper insight into the mutual influences between subsystems of language. The (in-/out-) degree and strength distributions of the word-level layers are overlapped due to the same word frequencies reflected from the same data source. Therefore, the standard approach to study the structure of linguistic networks showed no discrepancies among layers. However, the (in-/out-) selectivity values are potentially capable of quantifying differences, namely to show the potential of revealing the interplay among the layers.

The inter layer degree and strength correlations suggest that CO-SHU layers are more related than the CO-SIN, and SIN-SHU pairs, due to the preserving Zipf's law during shuffling [21] (reflecting the utilization of the same data source). In-distributions for syntax layers in both languages have higher values than the corresponding out-distributions, and generally SIN is less inter correlated than the CO and SHU layers. The inter and intra layer correlations in the multilayer language network suggest the manifestation of different governing principles in the syntax structure

of the examined languages. This is primarily reflection of the rich Croatian morphology. The interesting part is that this is the first observable indication of differences between languages manifested in a multilayer analysis framework, which encouraged a deeper investigation. In addition, the selectivity distributions (regardless of side or layer or language) are not correlated, supporting the potential of selectivity as a measure capable to quantify structural differences across language subsystems. Moreover, Croatian exhibits higher correlations then English in general.

The examination of the word-level layers overlap reveals additional insights into the mutual interplay between the layers. The weighted overlap provides a thorough insight into the intersection of links between network layers. It seems that WO is more appropriate to approximate the overlaps of layers in weighted networks than the commonly employed Jaccard measure. As expected, CO-SIN layers are more overlapped than shuffled pairs, and Croatian syntax is better captured through words co-occurrences than the English. The preserved weights on intersected links indicate that around 10% of the co-occurrence frequencies are not consistent with overlapped syntax dependencies. The proposed measure of preserved weighted overlap seems adequate to quantify the similarity of word-level layers in weighted and directed multilayer networks of language.

The subword layer's analysis reveals that the syllabic layer plays an important role in the manifestation of principles governing the construction of word layer, which is different for the examined languages. The graphemic layers, on the other hand, share characteristics, which are reflections of the high density of the graphemic networks (almost complete graphs in both languages).

The obtained multilayered language analysis results manifest different driving principles beneath the co-occurrence, shuffled, syntactic, syllabic and graphemic layers, which was not obvious through the analysis of isolated layers. In order to obtain deeper insight into these relations we utilize the analysis of motifs, which reveal a close topological structure in the syntactic and syllabic layers of both languages. The correlations of the motifs' frequencies are more emphasized in Croatian. The triad significance profiles (TSP) are correlated between syntax and syllables regardless of the language, while English additionally exhibits a correlation between co-occurrence and syntax layers. It seems that the observed TSP correlations reflect the properties of the Croatian - the free word-order which caused different characterizations of the co-occurrence and syntax layers. Moreover, the high flectivity of Croatian is reflected in many suffixes realized by syllables. Therefore, the structure of layers also reflects the morphological properties inherent to the language, which should we examine more deeply in the future.

Our findings are in line with previous observations in language networks research. For instance, Ferrer i Cancho [60] reports that the amount of syntactically incorrect links in co-occurrence networks can increase to a high of 70%, and elaborates: "About 90% of syntactic relationships take place at a distance lower or equal than two, but word co-occurrence networks lack a linguistically precise definition of link and fail in capturing the characteristic long-distance correlations of words in sentences." This adequately explains the driving principle of the CO-SIN relationships which we have confirmed in this research. Still, an explanation of the linguistic grounding for the SIN-SYL relationships remains an open challenge.

Our results strongly suggest that there are some properties which are inherent in the word-level layers and not for the subword layers; while some are inherent in the word-subword relations. More precisely, it seems that syntax and syllables exhibit influences of the same linguistic phenomena.

16.6 Conclusion

In this research we use the multilayer networks' framework to explore various language subsystems' interactions. Multilayer networks are constructed from five variations of the same original text: three on the word-level (syntax, co-occurrence and it's shuffled counterpart) and two on the subword-level (syllables and graphemes). The analysis and comparison of layers at word and subword-levels is

employed in order to determine the mechanism of mutual interactions between different linguistic units.

The presented findings corroborate that the multilayer framework can meet the demands in expressing the complex structure of language. According to these results one can notice substantial differences between the networks' structures of different language layers, which are hidden during the exploration of an isolated layer, regardless of modeled language (e.g. Croatian or English). Therefore, it is important to include all language layers simultaneously in order to capture all language characteristics in the systematic exploration.

The multilayer network framework is a powerful, consistent and systematic approach to model several linguistic subsystems simultaneously and to provide a more general view on language. The word-level layers can be represented as multiplex networks (the coupled links have 1:1 or 0:1 inter-connections), while the connections between word and subword layers are not coupled (have N:M inter-connections). Hence, defining the unified theoretical model for the multilayer language networks is essential for further endeavors in the research of linguistic networks.

These findings reveal a variety of new and thrilling questions which will open new paths for future research in network linguistics. To conclude, we are at the very beginning of an exciting and challenging pursuit. Hence, our future research plans involve: exploring the relationships of other languages' subsystems (i.e. morphological, phonetic), defining the theoretical model capable of capturing all structural variations of language subsystems' relationships and eventually explain the governing principle of mutual interactions and conceptual universalities in natural languages.

IV

Bibliography

Bibliography

[1] Alstott, J., Bullmore, E. and Plenz, D. Powerlaw: a python package for analysis of heavy-tailed distributions. *arXiv preprint arXiv:1305.0215*, 2013.

[2] Biemann, C., Roos, S. and Weihe K. Quantifying semantics using complex network analysis, Proceedings of the 24th International Conference on Computational Linguistics-COLING 2012, 263-278, 2012.

[3] Borge-Holthoefer, J. and Arenas, A. Semantic networks: Structure and dynamics. Entropy, 12(5):1264-1302, 2010.

[4] Caldeira, S., Lobao,P., Andrade, R., Neme, A. and Miranda, V. The network of concepts in written texts. The European Physical Journal B-Condensed Matter and Complex Systems, 49(4), 523-529, 2006.

[5] Choudhury, M., Chatterjee, D. and Mukherjee, A. Global topology of word co-occurrence networks: Beyond the two-regime power-law. In Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, 162-170, 2010.

[6] Choudhury, M. and Mukherjee, A. The structure and dynamics of linguistic networks, Dynamics on and of complex networks, Modeling and Simulation in Science, Engineering and Technology, Springer, 145-166, 2009.

[7] Dorogovtsev, S. N. and Mendes, J. F. F. Language as an evolving word web. Proceedings of the Royal Society of London. Series B: Biological Sciences, 268(1485):2603-2606, 2001.

[8] Hagberg, A. A., Schult, D. A. and Swart P. J. Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors, Proc. 7th Python in Sci. Conf., Pasadena, CA USA: SciPy, 11–16, 2008.

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

[10] Liu, H. and Cong, J. Language clustering with word co-occurrence networks based on parallel texts. Chinese Science Bulletin, 58(10), 1139-1144, 2013.

[11] Masucci, A. P. and Rodgers, G. J. Network properties of written human language. Physical Review E, 74(2), 026102, 2006.

[12] Pardo, T. A. S., Antiqueira, L., das Gracas Nunes, M., Oliveira, O. N. and da Fontoura Costa, L. Using complex networks for language processing: The case of summary evaluation. In Communications, Circuits and Systems Proceedings, IEEE, vol. 4, 2678-2682. 2006.

[13] Newman, M. E. J. The structure and function of complex networks. SIAM review, 45(2), 167-256, 2003.

[14] Ban, K., Martinčić-Ipšić, S. and Meštrović, A. Initial comparison of linguistic networks measures for parallel texts. 5th International Conference on Information Technologies and Information Society (ITIS), 97-104, 2013.

[15] Krishna,M., Hassan, A., Liu, Y. and Radev, D. The effect of linguistic constraints on the large scale organization of language. arXiv preprint arXiv:1102.2831, 2011.

[16] Li, W. Random texts exhibit Zipf's-law-like word frequency distribution. IEEE Transactions on Information Theory, 38(6), 1842-1845, 1992.

[17] Liu, H. and Hu, F. What role does syntax play in a language network?. Europhysics Letters, 83(1), 18002, 2008.

[18] Margan, D., Martinčić-Ipšić, S. and Meštrović, A. Preliminary report on the structure of Croatian linguistic co-occurrence networks. 5th International Conference on Information Technologies and Information Society (ITIS), 89-96, 2013.

[19] Masucci, A. P. and Rodgers, G. J. Differences between normal and shuffled texts: structural properties of weighted networks. Advances in Complex Systems, 12(01), 113-129, 2009.

[20] Watts, D. J. and Strogatz, S. H. Collective dynamics of small-world networks. Nature, 393(6684), 440-442, 1998.

[21] Margan, D., Martinčić-Ipšić, S. and Meštrović, A. Network Differences Between Normal and Shuffled Texts: Case of Croatian. Studies in Computational Intelligence, Complex Networks V, vol. 549, 275-283, 2014.

[22] Solé, R. V., Corominas-Murtra, B., Valverde, S. and Steels, L. Language Networks: their structure, function and evolution, Complexity, 15(6), 20–26. doi:10.1002/cplx.20305, 2010.

[23] Ferrer i Cancho, R., Solé, R. V. and Köhler, R. Patterns in syntactic dependency networks, Physical Review E, 69(5), 51915, 2004.

[24] Liu, H. and Chunshan, X. Can syntactic networks indicate morphological complexity of a language?, Europhysics Letters, 93(2), 28005, 2011.

[25] Arbesman, S., Strogatz, S. H. and Vitevitch, M. S. Comparative analysis of networks of phonologically similar words in English and Spanish, Entropy, 12(3), 327-337, 2010.

[26] Arbesman, S., Strogatz, S. H. and Vitevitch, M. S. The Structure of Phonological Networks across Multiple Languages, International Journal of Bifurcation and Chaos, 20(3), 679-685, 2010.

[27] Pembe, F. C. and Bingol, H. Complex Networks in Different Languages: A Study of an Emergent Multilingual Encyclopedia, Unifying Themes in Complex Systems: Proceedings of the Sixth International Conference on Complex Systems. Springer, Berlin, Heidelberg, 612–617, 2010.

[28] Sheng, L. and Li, C. English and Chinese language as weighted networks, Physica A, 388, 2561-2570, 2009.

[29] Bastian, M., Heymann, S. and Jacomy, M. Gephi: an open source software for exploring and manipulating networks, In Proceedings of the 8th International AAAI Conference on Web and Social Media, 361-362, 2009.

[30] Medeiros Soares, M., Corso, G. and Lucena, L. S. The network of syllables in Portuguese, Physica A: Statistical Mechanics and its Applications, 355(2), 678-684, 2005.

[31] Ban, K., and Ivakić, I. and Meštrović, A. A preliminary study of Croatian language syllable networks, Information and Communication Technology Electronics & Microelectronics (MIPRO), IEEE, 1296–1300, 2013.

[32] Mehri, A., Darooneh, A. H. and Shariati, A. The complex networks approach for authorship attribution of books, Physica A, 391(7), 2429-2437, 2012.

[33] Bird, S., Klein, E. and Loper, E. Natural Language Processing with Python, O'Reilly Media, 2009.

[34] The Center for Information and Language Processing, TreeTagger - a language independent part-of-speech tagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/, downloaded: June, 2013.

[35] Tadić, M. and Fulgosi, S. Building the Croatian Morphological Lexicon, Proceedings of the EACL2003, 41-46, 2003.

[36] Antiqueira, L., Pardo, T. A. S., das Gracas Volpe Nunes, M., Oliveira, O. N. Jr. Some issues on complex networks for author characterization, Inteligencia Artificial, 11(36), 51-58, 2007.

[37] Antiqueira, L., Oliveira, O. N. Jr, and da Fontoura Costa, L. and Nunes das Graças Volpe, M. A complex network approach to text summarization, Information Sciences, 179(5), 584–599, 2009.

[38] Antiqueira, L. et al. Strong correlations between text quality and complex networks features. Physica A: Statistical Mechanics and its Applications 373, 811-820, 2007.

[39] Erdös, P. and Rényi, A. On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 5(1), 17-60, 1960.

[40] Strogatz, S. H. Exploring complex networks. Nature, 410(6825), 268-276, 2001.

[41] Milgram, S. The Small World, Psychology Today, 1(1), 60-67, 1967.

[42] Albert, R., Barabási, A. L. Statistical mechanics of complex networks. Rev. Mod. Physics, 74, 47–97. 2002.

[43] Barabási, A. L. and Albert, R. Emergence of scaling in random networks, Science, 286(5439), 509-512, 1999.

[44] Peng, G., Minett, J. W. and Wang, W. S-Y. The network of Syllables and Characters in Chinese, Journal of Quantitative Linguistics, 15(3), 243-255, 2008.

[45] Bollobas, B. Random Graphs. Academic Press, London, 1985.

[46] Turk, M. Fonologija hrvatskoga jezika, Izdavački centar Rijeka, Tiskara Varaždin, Varaždin, 1992.

[47] Lewis, E. and Tatham, M. Word and syllable concatenation in text-to-speech synthesis, In:Sixth European Conference on Speech Communications and Technology, 615-618, 1999.

[48] Larson, M. and and Eickeler, S. Using syllable-based indexing features and language models to improve German spoken document retrieval, In Eurospeech'03, 1217-1220, 2003.

[49] Majewski, M. Syllable based language model for large vocabulary continuous speech recognition of Polish, In Text, Speech and Dialogue, 397-401, 2008.

[50] Margan, D., Meštrović, A., Martinčić-Ipšić, S. Complex Networks Measures for Differentiation between Normal and Shuffled Croatian Texts. In 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2014), IEEE, 1819-1823, 2014

[51] Milo, R., Shen-Orr, S. Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. Network motifs: simple building blocks of complex networks, Science, 298(5594), 824-827, 2002.

[52] Milo, R. et al. Superfamilies of evolved and designed networks, Science, 303(5663), 1538-1542, 2004.

[53] Pržulj, N. Biological network comparison using graphlet degree distribution. Bioinformatics 23(2), e177-e183. 2007.

[54] Rasche, F. and Wernicke, S. FANMOD fast network motif detection - manual, Bioinformatics, 22(9), 1152-1153, 2006.

[55] Wernicke, S. A faster algorithm for detecting network motifs. In R. Cassadio and G. Myers, editors, Proceedings of WABI'05, Lecture Notes in Computer Science, Springer, 3692, 165-177, 2005.

[56] Abramov, O. and Mehler, A. Automatic language classification by means of syntactic dependency networks, Journal of Quantitative Linguistics, vol. 18(4), 291-336, 2011.

[57] Bader, D. A. and Madduri, K. SNAP, Small-world Network Analysis and Partitioning: an open-source parallel graph framework for the exploration of large-scale networks, In Proceedings of the International Symposium on Parallel and Distributed Processing, IEEE, 1-12, 2008.

[58] Beliga, S. and Martinčić-Ipšić, S. Node selectivity as a measure for graph-based keyword extraction in Croatian news, In Proceedings of the 6th International Conference on Information Technologies and Information Society, Slovenia, 8-17, 2014.

[59] Beliga, S., Meštrović, A. and Martinčić-Ipšić, S. Toward selectivity based keyword extraction for Croatian news, In Proceedings of the Workshop on Surfacing the Deep and the Social Web, CEUR, Vol. 1301, 1-14, 2014.

[60] Ferrer i Cancho, R. The structure of syntactic dependency networks: insights from recent advances in network theory, Problems of Quantitative Linguistics, Studies in Computational Intelligence, 60-75, 2005.

[61] Ferrer i Cancho, R., Mehler, A., Pustylnikov, O. and Diaz-Guilera, A. Correlations in the Organization of Large-Scale Syntactic Dependency Networks, Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing, ACM, 65-72, 2007.

[62] Csardi, G. and Nepusz, T. The igraph software package for complex network research, Inter. Journal Complex Systems, 1695(5), 1-9, 2006.

[63] Hansen, D., Shneiderman, B. and Smith, M. A. Analyzing social media networks with NodeXL: Insights from a connected world, Morgan Kaufmann, 2010.

[64] Kuchaiev, O., Stevanovć, A., Hayes, W. and Pržulj, N. GraphCrunch 2: software tool for network modeling, alignment and clustering, Bioinformatics, 12(1), 24, 2011.

[65] Margan, D., Meštrović, A., Ivašić-Kos, M. and Martinčić-Ipšić, S. Toward a Complex Networks Approach on Text Type Classification, 6th International Conference on Information Technologies and Information Society, Slovenia, 2014.

[66] Milenković, T. and Pržulj, N. Uncovering biological network function via graphlet degree signatures, Cancer Informatics, 6, 257–273, 2008.

[67] Newman, M. E. J. Networks: An Introduction, Oxford University Press, 2010.

[68] De Nooy, W., Mrvar, A. and Batagelj, V. Exploratory social network analysis with Pajek. Cambridge University Press, Vol. 27, 2011.

[69] Rizvić, H., Martinčić-Ipšić, S. and Meštrović, A. Network motifs analysis of Croatian literature, In Proceedings of the 6th International Conference on Information Technologies and Information Society, Slovenia, 2-7, 2014.

[70] Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Research, 13(11), 2498-2504, 2003.

[71] Šišović, S., Martinčić-Ipšić, S. and Meštrović, A. Comparison of the language networks from literature and blogs, In Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, IEEE, Croatia, 1824-1829, 2014.

[72] Šišović, S., Martinčić-Ipšić, S. and Meštrović, A. Toward network-based keyword extraction from multitopic web documents, In Proceedings of the 6th International Conference on Information Technologies and Information Society, Slovenia, 18-27, 2014.

[73] Python Software Foundation. Unicode Database Documentation, [Online]. Available: https://docs.python.org/2/library/unicodedata.html#unicodedata.normalize [Accessed: April 10, 2015].

[74] The Unicode Consortium. Unicode Standard Annex #15: Unicode normalization forms, http://unicode.org/reports/tr15/#Norm_Forms [Accessed: April 10, 2015].

[75] Abilhoa, W. D., de Castro, L. N. A keyword extraction method from twitter messages represented as graphs. Applied Mathematics and Computation, 240, 308-325, 2014.

[76] Ahel, R., Dalbelo-Bašić, B. and Šnajder, J. Automatic keyphrase extraction from Croatian newspaper articles. In Proceesings of The Future of Information Sciences, Digital Resources and Knowledge Sharing, 207-218, 2009.

[77] Bekavac, M. and Šnajder, J. GPKEX: Genetically Programmed Keyphrase Extraction from Croatian Texts. In Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, ACL, 43-47, 2013.

[78] Berry, M. W. Castellanos, M. Survey of Text Mining II, Springer, 2008.

[79] Berry, M. W. and Kogan, J. Text Mining: Applications and Theory, Wiley, 2010.

[80] Boudin, F. A comparison of centrality measures for graph-based keyphrase extraction. In International Joint Conference on Natural Language Processing (IJCNLP), 834-838, 2013.

[81] Chang, J.-Y. Kim, I.-M. Analysis and Evaluation of Current Graph-Based Text Mining Researches. Advanced Science and Technology Letters, Mobile and Wireless 2013, 42, 100-103, 2013.

[82] Chen, P. and Lin, S. Automatic keyword prediction using Google similarity distance. Expert Systems with Applications, 37(3), 1928-1938, 2010.

[83] Divjak, B. and Lovrenčić, A. Diskretna matematika s teorijom grafova. TIVA Tiskara Varaždin, 2005.

[84] Dobša, J. Information retrieval using latent semantic indexing. Journal of Information and Organizational Sciences, 26(1-2), 13-23, 2002.

[85] Erkan, G. and Radev, D. LexRank: Graph-based lexical centrality as salience in text summarization. Journal od Articial Intelligence Research (JAIR), 22(1), 457-479, 2004.

[86] Feldman, R. and Sanger, J. The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press, 2007.

[87] Grineva, M., Grinev, M. and Lizorkin, D. Extracting Key Terms from Noisy and Multi-theme Documents. In Proceedings of the 18th International Conference on World Wide Web, ACM, 661-670, 2009.

[88] HaCohen-Kernen, Y. and Gross, Z. Masa, A. Automatic Extraction and Learning of Keyphrases from Scientific Articles. Computational Linguistics and Intelligent Text Processing. In Proceedings of 6th Int. Conference CICLing 2005, LNCS 3406, 657-669, 2005.

[89] HaCohen-Kerner, Y. Automatic Extraction of Keywords from Abstracts. Knowledge-Based Intelligent Information and Engineering Systems, In Proceedings of 7th International Conference KES 2003, (LNCS 2773), 843-849, 2003.

[90] Hotho, A., Nürnberger, A. and Paass, G. A Brief Survey of Text Mining. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 20(1), 19-62, 2005.

[91] Huang, C., Tian, Y., Zhou, Z., Ling, C. X. and Huang, T. Keyphrase extraction using semantic networks structure analysis. In IEEE International Conference on Data Mining, ICDM'06, 275-284, 2006.

[92] Hulth, A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In Proceedings of EMNLP 2003, 216-223, 2003.

[93] Hurt, C. D. Automatically Generated Keywords: A Comparison to Author-Generated Keywords in the Sciences. Journal of Information and Organizational Sciences, 34(1), 81-88, 2010.

[94] Jones, K. S. Information retrieval and artificial intelligence. Artificial Intelligence, 114(1-2), 257-281, 1999.

[95] Krapivin, M., Autayeu, A., Marchese, M., Blanzieri, E. and Segata, N. Keyphrases Extraction from Scientific Documents: Improving Machine Learning Approaches with Natural Language Processing. In Proceedings of 12th International Conference on Asia-Pacific Digital Libraries, ICADL 2010, LNAI 6102, 102-111, 2010.

[96] Lahiri, S., Choudhury, S. R., and Caragea, C. Keyword and keyphrase extraction using centrality measures on collocation networks. arXiv preprint arXiv:1401.6571. 2014.

[97] Langville, A. and Meyer, C. A survey of eigenvector methods of web information retrieval. SIAM Review, 47(1):135-161, 2005.

[98] Litvak, M. and Last, M. Graph-based keyword extraction for single-document summarization. In Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, MMIES '08, 17-24, 2008.

[99] Litvak, M., Last, M., Aizenman, H., Gobits, I., and Kandel, A. DegExt-A language-independent graph-based keyphrase extractor. In Advances in Intelligent Web Mastering-3, 121-130, Springer Berlin Heidelberg. 2011.

[100] Manyika, J., Chui, J., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A. H. Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011.

[101] Matsuo, Y., Ohsawa, Y., Ishizuka, M. Keyworld: Extracting keywords from document s small world. In Discovery Science, 271-281, 2001.

[102] Medelyan, O. and Witten, I. H. Thesaurus Based Automatic Keyphrase Indexing. In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, 296-297, 2006.

[103] Meštrović, A., Martinčić-Ipšić, S. and Čubrilo, M. Weather forecast data semantic analysis in f-logic. Journal of Information and Organizational Sciences, 31(1), 115-129, 2007.

[104] Mihalcea, R. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics, ACL 2004, 20, 2004.

[105] Mihalcea, R. and Radev, D. Graph-based Natural Language Processing and Information Retrieval, Cambridge University Press, 2011.

[106] Mihalcea, R. and Tarau, P. TextRank: Bringing order into texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), ACL, 104-411, 2004.

[107] Mijić, J., Dalbelo-Bašić, B. and Šnajder, J. Robust Keyphrase Extraction for a Large-Scale Croatian News Production System. In Proceedings of International Conference on Formal Approaches to South Slavic and Balkan Languages, 59-66, 2010.

[108] Nguyen, T. D. and Kan, M.-Y. Keyphrase extraction in scientific publications. In Proceedings of ICADL 2007, 317-326, 2007.

[109] Ohsawa, Y., Benson, N. E. and Yachida, M. KeyGraph: Automatic Indexing by Co-Occurrence Graph Based on Building Construction Metaphor. In Proceedings of the Advances in Digital Libraries Conference, ADL '98, 12, 1998.

[110] Palshikar, G. K. Keyword Extraction from a Single Document Using Centrality Measures. Pattern Recognition and Machine Intelligence, LNCS 4815, 503-510, 2007.

[111] Paralič, J. and Marek, P. Some approaches to text mining and their potential for semantic web applications. Journal of Information and Organizational Sciences, 31(1), 157-169, 2007.

[112] Pasquier, C. Single Document Keyphrase Extraction Using Sentence Clustering and Latent Dirichlet Allocation. In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, 154-157, 2010.

[113] Pudota, N., Dattolo, A., Baruzzo, A. and Tasso, C. A New Domain Independent Keyphrase Extraction System. Digital Libraries, CCIS 2010, 91, 67-78, 2010.

[114] Saratlija, J., Šnajder, J. and Dalbelo-Bašić, B. Unsupervised topic-oriented keyphrase extraction and its application to Croatian. Text, Speech and Dialogue, LNCS 6836, 340-347, 2011.

[115] Sebastiani, F. Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1-47, 2002.

[116] Sonawane, S. S. and Kulkarni, P. A. Graph based Representation and Analysis of Text Document: A Survey of Techniques. International Journal of Computer Applications, 96(19):1-8, 2014.

[117] Song, M., Song, I.-Y. and Hu, X. KPSpotter: a flexible information gain-based keyphrase extraction system. In Proceedings of 5th International Workshop of WIDM 2003, 50-53, 2003.

[118] Tsatsaronis, G., Varlamis, I. and Nørvåg, K. SemanticRank: ranking keywords and sentences using semantic graphs. In Proceedings of the 23rd International Conference on Computational Linguistics, 1074-1082, 2010.

[119] Turney, P. D. Coherent Keyphrase Extraction via Web Mining. In Proceedings of the 18th International Joint Conference on AI, IJCAI'03, 434-439, 2003.

[120] Turney, P. D. Learning algorithms for keyphrase extraction. 2(4), 303–336, 2000.

[121] Turney, P. D. Learning to Extract Keyphrases from Text. In Technical Report ERB-1057, National Research Council of Canada, Institute for Information Technology, 1999.

[122] Wan, X. and Xiao, J. Single Document Keyphrase Extraction Using Neighborhood Knowledge. In Proceedings of the 23rd AAAI Conference on Artificial Intelligence, 855-860, 2008.

[123] Wang, J., Peng, H. and Hu, J.-S. Automatic Keyphrases Extraction from Document Using Neural Network. In Advances in Machine Learning and Cybernetics, 4th International Conference ICMLC 2005, LNCS 3930, 633-641, 2006.

[124] Washio, T. and Motoda, H. State of the Art of Graph-based Data Mining. SIGKDD Explorations Newsletter, 5(1), 59-68, 2003.

[125] Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C. and Nevill-Manning, C. G. Kea: Practical Automatic Keyphrase Extraction. In Proceedings of the 4th ACM Conference of the Digital Libraries, DL '99, 254-255, 1999.

[126] Wu, J-L. and Agogino, A. M. Automating Keyphrase Extraction with Multi-Objective Genetic Algorithms, In Proceedings of the 37th HICSS, IEEE, 4(4), 104-111, 2003.

[127] Xie, Z. Centrality Measures in Text Mining: Prediction of Noun Phrases that Appear in Abstracts. In Proceedings of the ACL Student Research Workshop, ACLstudent '05, 103-108, 2005.

[128] Yang, Z., Lei, J., Fan, K. and Lai, Y. Keyword extraction by entropy difference between the intrinsic and extrinsic mode. Physica A: Statistical Mechanics and its Applications, 392(19), 4523-4531, 2013.

[129] Zhang, C., Wang, H., Liu, Y., Wu, D., Liao, Y. and Wang, B. Automatic Keyword Extraction from Documents Using Conditional Random Fields. In Journal of Computational Information Systems, 4(3), 1169-1180, 2008.

[130] Zhang, K., Xu, H., Tang, J. and Li, J. Keyword Extraction Using Support Vector Machine. Advances in Web-Age Information Management, LNCS 4016, 85-96, 2006.

[131] Zhang, Y., Milios, E. and Zincir-Heywood, N. A Comparison of Keyword- and Keyterm-based Methods for Automatic Web Site Summarization. In Technical Report: Papers from the AAAI'04 Adaptive Text Extraction and Mining, 15-20, 2004.

[132] Zhou, Z., Zou, X., Lv, X. and Hu, J. Research on Weighted Complex Network Based Keywords Extraction. Chinese Lexical Semantics, LNCS 8229, 442-452, 2013.

[133] Liu, Z., Li, P., Zheng, Y. and Sun, M. Clustering to find exemplar terms for keyphrase extraction. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 1(1), 257-266. 2009.

[134] Meštrović, A. and Čubrilo, M. Monolingual dictionary semantic capturing using concept lattice. International Review on Computers and Software, 6(2), 173–184, 2011.

[135] Meštrović, A. Semantic Matching Using Concept Lattice. In Proc. of Concept Discovery in Unstructured Data, Katholieke Universiteit Leuven, 49–58, 2012.

[136] Opsahl, T. Agneessens, F. and Skvoretz, J. Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks, 32(3), 245–251, 2010.

[137] Tomokiyo, T. and Hurst, M. A language model approach to keyphrase extraction. In Proc. of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment, 18, 33–40, 2003.

[138] Amancio, D. R. et al. Complex networks analysis of language complexity, Europhysics Letters, 100(5), 58002, 2012.

[139] Amancio, D. R. et al. Using metrics from complex networks to evaluate machine translation, Physica A: Statistical Mechanics and its Applications, 390(1), 131-142. 2009.

[140] Costa, L. da F. et. al. Characterization of complex networks: a survey of measurements, Advances in physics, 56(1), 167-242. 2007.

[141] Twitter, Wikipedia, the free encyclopedia, 20-Feb-2016. [Online]. Available: https://en.wikipedia.org/wiki/Twitter. [Accessed: 21-Feb-2016].

[142] Huberman, B. A., Romero, D. M. and Wu, F. Social networks that matter: Twitter under the microscope, arXiv:0812.1045 [physics], Dec. 2008.

[143] Cha, M., Haddadi, H., Benevenuto, F. and Gummadi P. K. Measuring user influence in Twitter: The million follower fallacy. In: 4th Int. AAAI Conf. on Weblogs and Social Media, AAAI Press. 10-17. 2010.

[144] Bollen, J., Mao, H. and Zeng, X.-J. Twitter mood predicts the stock market, Journal of Computational Science, vol. 2, no. 1, 1–8, 2011.

[145] Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I. M. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In: 4th Int. AAAI Conf. on Weblogs and Social Media, AAAI Press. 178–185. 2010.

[146] Boyd, D., Golder, S. and Lotan, G. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter, in 2010 43rd Hawaii International Conference on System Sciences (HICSS), 1–10. 2010.

[147] Mathioudakis, M. and Koudas N. TwitterMonitor: Trend detection over the Twitter stream. In: Proc. 2010 ACM SIGMOD Int. Conf. on Management of Data, ACM. 1155–1158. 2010.

[148] Go, A., Bhayani, R. and Huang, L. Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford, 1, 12–16, 2009.

[149] Pak, A. and Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining., in LREC 2010, 10, 1320–1326. 2010.

[150] Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A. and Montejo-Ráez, A. Sentiment analysis in Twitter. Natural Lang Engineering, 20, 1–28. 2012.

[151] Agarwal, A., Xie, B., Vovsha, I. Rambow, O. and Passonneau, R. Sentiment Analysis of Twitter Data, in Proceedings of the Workshop on Languages in Social Media, 30–38, 2011.

[152] Kouloumpis, E., Wilson, T. and Moore, J. Twitter Sentiment Analysis: The Good the Bad and the OMG!, in Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[153] Wang, X., Wei, F., Liu, X., Zhou, M. and Zhang, M. Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach, in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 1031–1040, 2011.

[154] Villazon-Terrazas, J., Aparicio, S. and Alvarez, G. Study on Twitter as a Complex Network, in The Third International Conference on Digital Enterprise and Information Systems (DEIS2015), 54. 2015.

[155] Aparicio, S., Villazón-Terrazas, J. and Álvarez, G. A Model for Scale-Free Networks: Application to Twitter, Entropy, 17(8), 5848–5867, 2015.

[156] Lü, L. and Zhou, T. Role of Weak Ties in Link Prediction of Complex Networks, EPL 89, 18001. 2010.

[157] De Sá, H.R. and Prudêncio, R.B. Supervised link prediction in weighted networks. In: 2011 Int. Joint Conf. on Neural Netw., IEEE. 2281-2288. 2011.

[158] Yang, Y., Lichtenwalter, R. N. and Chawla, N.V. Evaluating link prediction methods. Knowledge and Information Systems. 45(3), 751–782. 2015.

[159] Baccianella, S., Esuli, A. and Sebastiani, F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, In: LREC, 2200-2204. 2010.

[160] GET search/tweets | Twitter Developers. [Online]. Available: https://dev.twitter.com/rest/reference/get/search/tweets. [Accessed: 21-Feb-2016].

[161] Margan, D. and Meštrović, A. LaNCoA: a Python toolkit for language networks construction and analysis. In Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention, IEEE, 1628–1633, 2015.

[162] Lehman, H. C. The exponential increase in man's cultural output. Social Forces, 25, 281-290. 1947.

[163] Evans, J. A. and Foster, J. G. Metaknowledge. Science, 331, 721–725. 2011.

[164] Michel, J. B., Shen, Y. K., Presser Aiden, A., Veres, A., Gray, M. K. et al. Quantitative analysis of culture using millions of digitized books. Science, 331, 176–182. 2011.

[165] Lazer, D., Pentland, A., Adamic, L. A., Aral, S., Barabási, A. L., et al. Life in the network: the coming age of computational social science. Science, 323, 721–723. 2009.

[166] Castellano, C., Fortunato, S. and Loreto, V. Statistical physics of social dynamics. Rev Mod Phys, 81, 591–646.2009.

[167] Fortunato, S. Community detection in graphs. Physics Reports, 486(3), 75–174.2010.

[168] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y. and Porter, M. A. Multilayer networks, Journal of Complex Networks, 2(3), 203-271, 2014.

[169] Boccaletti, S., Bianconi, G., Criado, R., del Genio, C., Gómez-Gardeñes, J., et al. The structure and dynamics of multilayer networks. Physics Reports, 544(1), 1-122, 2014.

[170] Pastor-Satorras, R., Castellano, C., Van Mieghem, P., Vespignani, A. Epidemic processes in complex networks. Rev Modern Physics, 87, 925, 2015

[171] Wang, Z., Bauch, C,T,, Bhattacharyya, S., d'Onofrio, A., Manfredi, P., et al. Statistical physics of vaccination. Physics Reports, 664, 1–113, 2016.

[172] Perc, M., Jillian, J. J., Rand, D. G., Wang, Z., Boccaletti, S., et al. Statistical physics of human cooperation. Physics Reports, 687, 1-51, 2017

[173] Althouse, B. M., Scarpino, S. V., Meyers, L. A., Ayers, J. W., Bargsten, M., et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. EPJ Data Science, 4, 1–17. 2015.

[174] Bessi, A., Zollo, F., Del Vicario, M., Scala, A., Caldarelli, G., et al. Trend of narratives in the age of misinformation. PLOS ONE, 10, e0134641. 2015.

[175] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., et al. The spreading of misinformation online. Proc Natl Acad Sci USA, 113, 554–559. 2016.

[176] González, M. C., Hidalgo, C. A., Barabási, A. L. Understanding individual human mobility patterns. Nature, 453, 779–782.2008.

[177] Palchykov, V., Mitrović, M., Jo, H. H., Saramäki, J. and Pan, R. K. Inferring human mobility using communication patterns. Scientific Reports, 4: 6174. 2014.

[178] Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F., Vespignani, A. Characterizing and modeling the dynamics of online popularity. Phys Rev Lett, 105, 158701. 2010.

[179] Preis, T., Moat, H.S. and Stanley, H.E. Quantifying trading behavior in financial markets using Google Trends. Scientific Reports, 3, 1684. 2013.

[180] Curme, C., Preis, T., Stanley, H.E. and Moat, H.S. Quantifying the semantics of search behavior before stock market moves. Proc Natl Acad Sci USA, 111, 11600–11605, 2014.

[181] Preis, T., Reith, D. and Stanley, H. E. Complex dynamics of our economic life on different scales: insights from search engine query data. Phil Trans R Soc A, 368, 5707–5719. 2010.

[182] Chatterjee, A., Mitrović, M. and Fortunato, S. Universality in voting behavior: an empirical analysis. Scientific Reports, 3, 1049. 2013.

[183] Conover, M., Ratkiewicz, J., Francisco, M.R., Gonçalves, B., Menczer, F., et al. Political polarization on Twitter. In: 5th Int. AAAI Conf. Weblogs and Social Media, AAAI Press, 133. 89–96. 2011.

[184] Mitrović, M., Paltoglou, G., Tadić, B. Networks and emotion-driven user communities at popular blogs. Eur Phys J B, 77, 597–609. 2010.

[185] Mitrović, M., Paltoglou, G., Tadić, B. Quantitative analysis of bloggers' collective behavior powered by emotions. J Stat Mech, 2011, P02005. 2011.

[186] Vicient, C. and Moreno, A. Unsupervised topic discovery in micro-blogging networks. Expert Systems with Applications, 42, 6472–6485, 2015.

[187] Song, S., Meng, Y. and Sun, J. Detecting keyphrases in micro-blogging with graph modeling of information diffusion. In: Pham DN, Park SB, editors, PRICAI 2014: Trends in Artificial Intelligence, Cham: Springer. 26–38, 2014.

[188] Rowe, M., Stankovic, M. and Alani, H. Who will follow whom? exploiting semantics for link prediction in attention-information networks. 11th Int. Semantic Web Conf., Berlin, Heidelberg: Springer. 476–491, 2012.

[189] Hong, L., Dan, O. and Davison, B. D. Predicting popular messages in Twitter. In: Proc. 20th Int. Conf. Companion on World Wide Web, New York, NY, USA: ACM. 57–58. 2011.

[190] Lü, L. and Zhou, T. Link prediction in complex networks: A survey. Physica A, 390, 1150–1170, 2011.

[191] Curiskis, S. A., Osborn, T. R. and Kennedy, P. J. Link Prediction and Topological Feature Importance in Social Networks. Australian Comp. Soc. Inc. 39-50. 2015.

[192] Liben-Nowell, D. and Kleinberg, J. The link-prediction problem for social networks. J Amer Soc for Inf Sci and Tech, 58, 1019–1031. 2017.

[193] Al Hasan, M., Zaki, M. J. A survey of link prediction in social networks. Social network data analytics, Boston, MA, Springer USA. 243–275. 2011.

[194] Bliss, C. A., Frank, M. R., Danforth, C. M. and Dodds, P. S. An evolutionary algorithm approach to link prediction in dynamic social networks. J of Computat Sci, 5, 750–764. 2014.

[195] He, Y., Liu, J. N., Hu, Y. and Wang, X. OWA operator based link prediction ensemble for social network. Expert System with Applications, 42, 21–50. 2015.

[196] Murata, T. and Moriyasu, S. Link prediction of social networks based on weighted proximity measures. In: Proc. IEEE/WIC/ACM int. conf. on web intelligence, IEEE, ACM. 85–88. 2007.

[197] Zhao, J., Miao, L., Yang, J., Fang, H., Zhang, Q. M. et al. Prediction of links and weights in networks by reliable routes. Scientific Reports, 5, 12261. 2015.

[198] Sharma, S. and Singh, A. An efficient method for link prediction in weighted multiplex networks. Comput Social Netw, 3, 7, 2016.

[199] Adamic, L. A. and Adar, E. Friends and neighbors on the web. Social Netw,25, 211–230, 2003.

[200] Martinčić-Ipšić, S., Miličić, T. and Meštrović, A. Text type differentiation based on the structural properties of language networks. Int. Conf. on Infor. and Software Tech., Cham: Springer. 536–548. 2016.

[201] Beliga, S., Meštrović, A., and Martinčić-Ipšić, S. An overview of graph-based keyword extraction methods and approaches. Journal of Information and Organizational Sciences, 39(1), 1-20. 2015.

[202] Beliga, S., Meštrović, A., and Martinčić-Ipšić, S. Selectivity-Based Keyword Extraction Method. International Journal on Semantic Web and Information Systems (IJSWIS), 12(3), 1-26, 2016.

[203] Valverde-Rebaza, J. and de Andrade Lopes, A. Exploiting behaviors of communities of twitter users for link prediction. Soc Netw Analysis and Mining, 3, 1063–1074. 2013.

[204] Newman, M. E. Clustering and preferential attachment in growing networks. Phys Rev E, 64, 025102. 2001.

[205] Barabâsi, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., et al. Evolution of the social network of scientific collaborations. Physica A, 311, 590–614. 2002.

[206] Zhou, T., Lü, L. and Zhang, Y.C. Predicting missing links via local information. Eur Phys J B, 71, 623–630. 2009.

[207] Fawcett, T. An introduction to ROC analysis. Pattern Recognition Letters, 27, 861–874, 2006.

[208] Demšar, J. Statistical comparisons of classifiers over multiple data sets. J of Machine Learning Research, 7, 1–30. 2006.

[209] Martinčić-Ipšić, S,, Margan, D. and Meštrović, A. Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. Physica A, 457, 117–128. 2016.

[210] Cong, J. and Liu, H. Approaching human language with complex networks, Physics of life reviews, 11(4), 598-618, 2014.

[211] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10, P10008, 2008.

[212] Caldarelli, G., Capocci, A., Servedio, V., Buriol, L., Donato, D., and Leonardi, S. Preferential attachment in the growth of social networks: the case of Wikipedia. In APS Meeting Abstracts, 1, 33003. 2006.

[213] Fang, Z., Wang, J., Liu, B. and Gong, W. Wikipedia as domain knowledge networks: domain extraction and statistical measurement. In: Proceedings of international conference on knowledge discovery and information retrieval (KDIR 2011), 159-165, 2011.

[214] Masucci, A. P., Kalampokis, A., Eguíluz, V. M. and Hernández-García, E. Wikipedia Information Flow Analysis Reveals the Scale-Free Architecture of the Semantic Space. PLoS ONE, 6(2), e17333. doi:10.1371/journal.pone.0017333, 2011.

[215] Vivi, N. and Strube, M. Transforming Wikipedia into a large scale multilingual concept network. Artificial Intelligence, 194, 62-85. 2013.

[216] Zlatić, V., et al. Wikipedias: Collaborative web-based encyclopedias as complex networks. Physical Review E, 74(1), 016115. 2006.

[217] Zlatić, V. and Štefančić, H. Model of Wikipedia growth based on information exchange via reciprocal arcs. EPL (Europhysics Letters), 93(5), 58005, 2011.

[218] Scott, J. Social network analysis. Sage. 2012.

[219] Ya-Rui, Z. and Ding, M. Modeling the evolution of collaboration network and knowledge network and their effects on knowledge flow through social network analysis. Journal of Digital Information Management, 14(4), 2016.

[220] Easley, D. and Kleinberg, J. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press. 2010.

[221] Freeman, L. C. A set of measures of centrality based on betweenness. Sociometry, 35-41, 1977.

[222] Bonacich, P. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology, 2(1), 113-120, 1972.

[223] Estrada, E. Virtual identification of essential proteins within the protein interaction network of yeast. Proteomics, 6(1), 35-40, 2006.

[224] Estrada, E. and Hatano, N. Communicability in complex networks. Physical Review E, 77(3), 036111, 2008.

[225] Estrada, E., Higham, D. J. and Hatano, N. Communicability betweenness in complex networks. Physica A: Statistical Mechanics and its Applications, 388(5), 764-774, 2009.

[226] Hahn, M. W. and Kern, A. D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Molecular biology and evolution, 22(4), 803-806. 2005.

[227] Guimera, R., Mossa, S., Turtschi, A. and Amaral, L. N. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. Proceedings of the National Academy of Sciences, 102(22), 7794-7799. 2005.

[228] Holme, P. Congestion and centrality in traffic flow on complex networks. Advances in Complex Systems, 6(02), 163-176, 2003.

[229] Brandes, U. and Fleischer, D. Centrality measures based on current flow. In Annual Symposium on Theoretical Aspects of Computer Science. Springer Berlin Heidelberg, 533-544, 2005.

[230] Latora, V., and Marchiori, M. A measure of centrality based on network efficiency. New Journal of Physics, 9(6), 188, 2007.

[231] Jian, Y. Keyword Extraction From Chinese Text Based On Multidimensional Weighted Features. Journal of Digital Information Management, 14(3), 2016.

[232] Matas, N., Martinčić-Ipšić, S. and Meštrović, A. Extracting domain knowledge by complex networks analysis of Wikipedia entries. In Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 1622-1627, 2015.

[233] Wikipedia. Online, Cited: April 2017. https://en.wikipedia.org/wiki/Wikpedia, 2017.

[234] Wikipedia. Online, Cited: April 2017. https://en.wikipedia.org/wiki/Wikipedia:Wikipedians, 2017.

[235] Silva, F. N., Viana, M. P., Travençolo, B. A. N. and Costa, L. D. F. Investigating relationships within and between category networks in Wikipedia. Journal of informetrics, 5(3), 431-438. 2011.

[236] West, R., Precup, D. and Pineau, J. Completing wikipedia's hyperlink structure through dimensionality reduction. In Proceedings of the 18th ACM conference on Information and knowledge management, 1097-1106, 2009.

[237] Itakura, K. Y., Clarke, C. L., Geva, S., Trotman, A. and Huang, W. C. Topical and structural linkage in Wikipedia. In European Conference on Information Retrieval, Springer Berlin Heidelberg, 460-465, 2011.

[238] Bargigli, L., di Iasio, G., Infante, L., Lillo, F. and Pierobon, F. Interbank markets and multiplex networks: centrality measures and statistical null models. In Interconnected Networks, Springer International Publishing. 179-194, 2016.

[239] Richardson, L. Beautiful soup documentation. 2007.

[240] Wikipedia Miner. Online, Cited: April 2017. http://wikipedia-miner.cms.waikato.ac.nz/, 2017.

[241] Avrachenkov, K., Litvak, N., Medyanikov, V. and Sokol, M. Alpha current flow betweenness centrality. In International Workshop on Algorithms and Models for the Web-Graph. Springer, Cham. 106-117, 2013.

[242] Costa, L. et al. Analyzing and modeling real-world phenomena with complex networks: a survey of applications, Advances in Physics, 60(3), 329-412, 2011.

[243] Amancio, D. and Oliveira Jr, Osvaldo, N, and Costa, L. Structure–semantics interplay in complex networks and its effects on the predictability of similarity in texts, Physica A: Statistical Mechanics and its Applications, 391(18), 4406–4419, 2012.

[244] Amancio, D. and Oliveira Jr, Osvaldo N. and Costa, L. Unveiling the relationship between complex networks metrics and word senses, EPL (Europhysics Letters), 98(1),18002 2012.

[245] Radev, D. and Mihalcea, R. Networks and natural language processing, AI magazine, 3(29), 16-28, 2008.

[246] Kurant, M. and Thiran, P. Layered complex networks, Physical review letters, 96(13), 138701-138705, 2006.

[247] Berlingerio, M., Coscia M., Giannotti, F. Monreale, A. and Pedreschi, D. Foundations of multidimensional network analysis, Advances in Social Networks Analysis and Mining (ASONAM), 485–489, 2011.

[248] Cardillo, A., Gómez-Gardeñes, J. Zanin, M., Romance, M., Papo, D., del Pozo, F. and Boccaletti, S. Emergence of network features from multiplexity, Scientific reports, 3, 1344, 2013.

[249] Menichetti, G. Remondini, D., Panzarasa, P., Mondragón, R. J. and Bianconi, G. Weighted Multiplex Networks, PloS ONE, 9(6), e97857, 2014.

[250] Gao, J., Buldyrev, S. V., Stanley, H. E. and Havlin, S. Networks formed from interdependent networks, Nature physics, 8(1), 40-48, 2012.

[251] Gao, J., Li, D. and Havlin, S. From a single network to a network of networks, National Science Review, 1(3), 346–356, 2014.

[252] Estrada, E. and Gómez-Gardeñes, J. Communicability reveals a transition to coordinated behavior in multiplex networks, Phys. Rev. E, 89(4), 042819, 2014.

[253] Barigozzi, M., Fagiolo, G. and Garlaschelli, D. Multinetwork of international trade: A commodity-specific analysis, Physical Review E, 81(4), 046104, 2010.

[254] Szell, M., Lambiotte, R. and Thurner, S. Multirelational organization of large-scale social networks in an online world, Proceedings of the National Academy of Sciences, 107(31), 13636–13641, 2010.

[255] Bullmore, E. and Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems, Nature Reviews Neuroscience, 10(3), 18–198, 2009.

[256] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A. and Onnela, J. Community structure in time-dependent, multiscale, and multiplex networks, Science, 328(5980), 876–878, 2010.

[257] HOBS, Croatian corpus, ver. 1.0, University of Zagreb, Faculty of Humanities and Social Sciences, http://metashare.elda.org/repository/browse/croatian-dependency-treebank/, accessed [2015-01-20], 2013.

[258] Tadić, M. Building the Croatian dependency treebank: the initial stages, Suvremena Lingvistika, 63(1), 85-97, 2007.

[259] Bianconi, G., Dorogovtsev, S. N. and Mendes, J. F. F. Mutually connected component of network of networks, Phys. Rev. E., 91(1), 012804, 2015.

[260] Morris, R. G. and Barthelemy, M. Transport on coupled spatial networks, Phys. Rev. Lett., 109(12), 128703, 2012.

[261] De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y, Porter, M. A., Gómez, S. and Arenas, A. Mathematical formulation of multilayer networks, Phys. Rev. X., 3(4), 041022–37, 2013.

[262] Chomsky, N. Aspects of the Theory of Syntax, MIT press, 1988.

[263] Jackendoff, R. What is the human language faculty?: Two views, Language, 87(3), 586-624, 2011.

[264] Brighton, H., Smith, K. and Kirby, S. Language as an evolutionary system, Physics of Life Reviews, 2(3), 177-226, 2005.

[265] Steels, L. Modeling the cultural evolution of language, Physics of Life Reviews, 884, 339-356, 2011.

[266] Hickok, G. The functional neuroanatomy of language, Physics of Life Reviews, 683, 121-143, 2009.

[267] Fenk-Oczlon, G. and Fenk, A. Complexity trade-offs between the subsystems of language, Language Complexity: Typology, Contact, Change, John Benjamins Publishing, 43-65, 2008.

[268] Diego R. Amancio, D. R., Nunes, M. G. V., Oliveira, O. N. Jr. and da F. Costa, L. Extractive summarization using complex networks and syntactic dependency, Physica A: Statistical Mechanics and its Applications, 391(4), 1855, 1864, 2012.

[269] Feldman, J. Embodied language, best-fit analysis, and formal compositionality, Physics of Life Reviews, 7(4), 385-410, 2010.

[270] Lyon, C., Nehaniv, C. L. and Saunders, J. Interactive language learning by robots: the transition from babbling to word forms, PLoS ONE, 7(6), e38236, 2012.

[271] Oudeyer, P.Y. The origins of syllable systems: an operational model, Proceedings of the 3rd ICCS International Conference on Cognitive Science, COGSCI 2001, 27-31, 2001.

[272] Kello, C. T. and Beltz, B. C. Scale-free networks in phonological and orthographic wordform lexicons, Approaches to Phonological Complexity, Walter de Gruyter, 171-190, 2009.

[273] Vitevitch, M. S. What can graph theory tell us about word learning and lexical retrieval?, Jour. of Speech, Language, and Hearing Research, 51(2), 408-422, 2008.

[274] Gruenenfelder, T. M. and Pisoni, D. B. The lexical restructuring hypothesis and graph theoretic analyses of networks based on random lexicons, Journal of Speech, Language, and Hearing Research, 52(3), 596-609, 2009.

[275] Mukherjee, A., Choudhury, M., Basu, A. and Ganguly, N. Self-organization of the sound inventories: analysis and synthesis of the occurrence and co-occurrence networks of consonants, Journal of Quantitative Linguistics, 16(2), 157-184, 2009.

[276] Vitevitch, M. S., Chan, K. Y. and Goldstein, R. Insights into failed lexical retrieval from network science, Cognitive Psychology, 68, 1–32, 2014.

[277] Marcus, M. P., Marcinkiewicz, M. A. and Santorini, B. Building a large annotated corpus of English: the Penn treebank, Computational Linguistics, 19(2), 313-330, 1993.

[278] Penn Treebank, English corpus, University of Pennsylvania, Computer and Information Science Department, http://www.nltk.org/howto/corpus.html/parsed-corpora, accessed [2015-03-25], 1995.

[279] BEEP pronouncing dictionary - syllabified, C. Barker, http://semarch.linguistics.fas.nyu.edu/barker/Syllables/index.txt, accessed [2014-03-25], 2002.

[280] Meštrović, A., Martinčić-Ipšić, S. and Matešić, M. Syllabification Based on Maximal Onset Principle for Croatian / Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik, Govor / Speech, 32(1), 3-35, 2015.

[281] Liu, H. and Cong, J. Empirical characterization of modern Chinese as a multi-level system from the complex network approach, Journal of Chinese Linguistics, 42(1), 1–38, 2014.

[282] Ban Kirigin, T., Meštrović, A. and Martinčić-Ipšić, S. Towards a Formal Model of Language Networks, Information and Software Technologies, Communications in Computer and Information Science, Springer, vol. 538, 469-479, 2015.

Index

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W

Z

ISBN 978-953-7720-34-6