LaNCoA: A Python Toolkit for Language Networks Construction and Analysis

Domagoj Margan, Ana Meštrović
Department of Informatics,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
Email: {dmargan, amestrovic}@uniri.hr

***Abstract*—In this paper 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.**

I. 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 [31]. 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 [14], [23], [24], [28], [35] or syntactic networks [1], [11]–[14], [21], [22]. There are also experiments with shuffled (randomized) networks [10], [18], [21], [28], [29]. Furthermore, syllables networks [4], phoneme networks [2] or semantic networks [9] 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 [24]. There are various software, tools and packages designed and developed for the task of complex networks analysis: Gephi [6], NodeXL [17], SNAP [3], Cytoscape [34], NetworkX package [16] for Python, igraph package [15] 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 [19] for graphlet analysis, GRAAL [30] for graph and network alignment or FANMOD [38] 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 [7], [8], [37] or text type classification [27].

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, 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 paper is structured as follows. In Section II we describe the language networks. In Section III we give a short overview of the complex networks analysis task. Then we present the LaNCoA toolkit in Section IV and we describe LaNCoA toolkit applications in Section V. We give a conclusion remarks

This work has been supported in part by the University of Rijeka under the project number 13.13.2.2.07.

in Section VI.

II. 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 [31] a detailed overview of all network measures and formulas is given.

III. 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 [14]. 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 [4].

The properties of the co-occurrence networks are derived from the word order in texts [14], [23], [24], [28], [35]. 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 [1], [11]–[14], [21], [22]. 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 [10], [18], [21], [28], [29]. 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. [4]. 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.

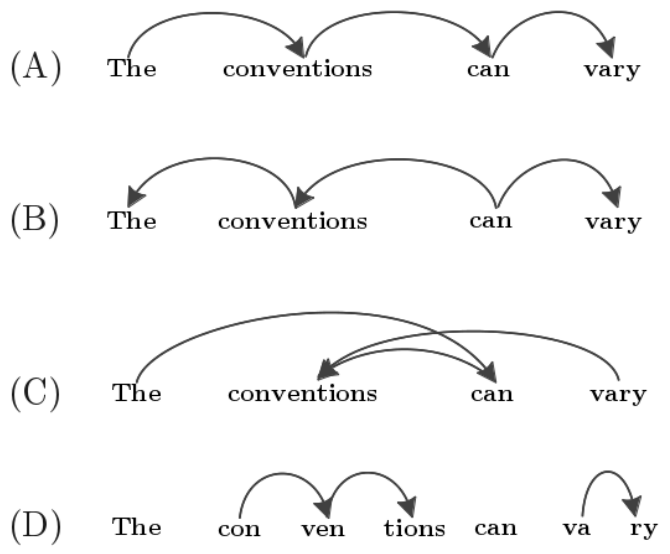

Fig. 1. Language networks construction rules presented on one toy-example sentence "The conventions can vary.": (A) Co-occurrence network, (B) Syntactic dependency network, (C) Shuffled network, (D) Syllable network

IV. 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 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 2. In this section we describe each module's function individually.

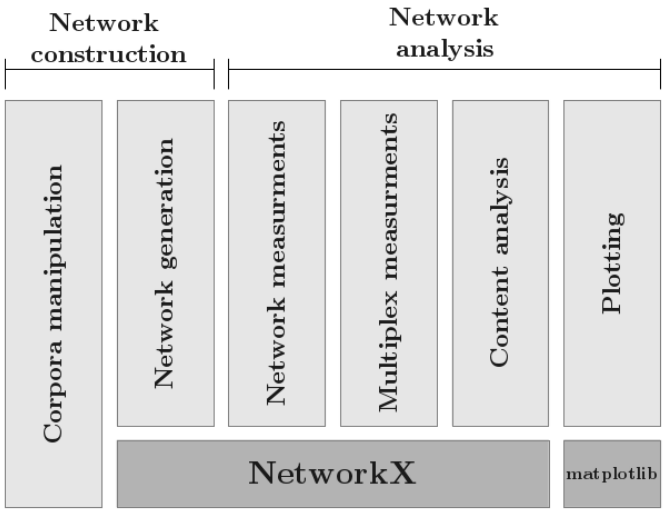

Fig. 2. LaNCoA architecture

A. Network Construction

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

1) 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:

a) 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 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 [39], [40] 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.

b) 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.

c) 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.

d) 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.

2) 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:

a) 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.

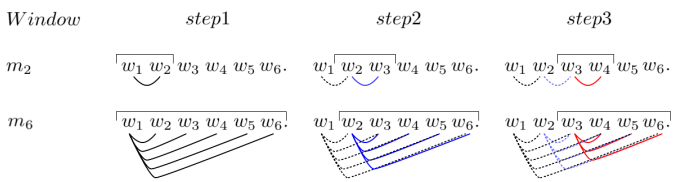

Fig. 3. An illustration of 3 steps in a network construction with a co-occurrence window m_n of sizes $n = 2$, and $n = 6$. $w_1 ... w_6$ are words within a sentence

b) 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.

c) 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.

d) 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.

B. 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.

1) 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.

2) 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.

3) 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.

4) 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.

V. The LaNCoA Toolkit Applications

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

In [24] 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 [5] 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 [25] 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 [26] 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 [36] we compared language networks from Croatian literature and blogs. In [7], [8], [37] 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 [33]. In [27] 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.

VI. Conclusion

In this paper 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.

References

[1] O. Abramov, and A. Mehler, "Automatic language classification by means of syntactic dependency networks," Journal of Quantitative Linguistics, vol. 18, no. 4, pp. 291-336, 2011.

[2] S. Arbesman, S.H. Strogatz, and M.S. Vitevitch, "The structure of phonological networks across multiple languages," Chaos, vol. 20, no. 3, pp. 679-685, 2010.

[3] D. A. Bader, and K. Madduri, "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, pp. 1-12, 2008.

[4] K. Ban, I. Ivakić, and A. Meštrović, "A preliminary study of Croatian language syllable networks," In Proceedings of the 36th International Convention on Information and Communication Technology, Electronics and Microelectronics, IEEE, Croatia, pp. 1296-1300, 2013.

[5] K. Ban, S. Martinčić-Ipšić, and A. Meštrović, "Initial comparison of linguistic networks measures for parallel texts," In Proceedings of the 5th International Conference on Information Technologies and Information Society, Slovenia, pp. 97-104, 2013.

[6] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," In Proceedings of the 8th International AAAI Conference on Web and Social Media, pp. 361-362, 2009.

[7] S. Beliga, and S. Martinčić-Ipšić, "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, 2014, pp. 8-17.

[8] S. Beliga, A. Meštrović and S. Martinčić-Ipšić, "Toward selectivity based keyword extraction for Croatian news," In Proceedings of the Workshop on Surfacing the Deep and the Social Web, CEUR, Vol. 1301, Italy, pp. 1-14, 2014.

[9] J. Borge-Hothoefer, and A. Arenas, "Semantic networks: structure and dynamics," Entropy, vol. 12, pp. 1264-1302, 2010.

[10] S. Caldeira, P. Lobao, R. Andrade, A. Neme, and V. Miranda, "The network of concepts in written texts," The European Physical Journal B - Condensed Matter and Complex Systems, vol. 49, no. 4, pp. 523-529, 2006.

[11] R. F. i Cancho, V. Sol, and R. Khler, "Patterns in syntactic dependency networks," Physical Review E, vol. 69, no. 5, 2004.

[12] R. F. i Cancho, "The structure of syntactic dependency networks: insights from recent advances in network theory," Problems of Quantitative Linguistics, pp. 60-75, 2005.

[13] R. F. i Cancho, A. Mehler, O. Pustylnikov, and A. Daz-Guilera, "Correlations in the organization of large-scale syntactic dependency networks," TextGraphs-2: Graph-Based Algorithms for Natural Language Processing, pp. 65-72, 2007.

[14] M. Choudhury, and A. Mukherjee, "The structure and dynamics of linguistic networks," Dynamics on and of Complex Networks, pp. 145-166, 2009.

[15] G. Csardi, and T. Nepusz, "The igraph software package for complex network research," InterJournal Complex Systems, vol. 1695, no. 5, pp. 1-9, 2006.

[16] A. Hagberg, P. Swart, and D. Chult, Exploring network structure, dynamics, and function using networkx, Technical report, Los Alamos National Laboratory (LANL), 2008.

[17] D. Hansen, B. Shneiderman, and M. A. Smith, Analyzing social media networks with NodeXL: Insights from a connected world, Morgan Kaufmann, 2010.

[18] M. Krishna, A. Hassan, Y. Liu, and D. Radev, "The effect of linguistic constraints on the large scale organization of language," arXiv preprint arXiv:1102.2831, 2011.

[19] O. Kuchaiev, A. Stevanovi, W. Hayes, and N. Pržulj, "GraphCrunch 2: software tool for network modeling, alignment and clustering, " Bioinformatics, vol. 12, no. 1, pp. 24, 2011.

[20] W. Li, "Random texts exhibit zipf's-law-like word frequency distribution," IEEE Transactions on Information Theory, vol. 38, no. 6, pp. 1842-1845, 1992.

[21] H. Liu, and F. Hu, "What role does syntax play in a language network?," Europhysics Letters, vol. 83, no. 1, pp. 18002, 2008.

[22] H. Liu, and C. Xu, "Can syntactic networks indicate morphological complexity of a language?," Europhysics Letters, vol. 93, no. 2, pp. 28005, 2011.

[23] H. Liu, and C. Jin, "Language clustering with word co-occurrence networks based on parallel texts," Chinese Science Bulletin, vol. 58, no. 10, pp. 1139-1144, 2013.

[24] D. Margan, S. Martinčić-Ipšić, and A. Meštrović, "Preliminary report on the structure of Croatian linguistic co-occurrence networks," In Proceedings of the 5th International Conference on Information Technologies and Information Society, Slovenia, pp. 89-96, 2013.

[25] D. Margan, S. Martinčić-Ipšić and A. Meštrović, "Network differences between normal and shuffled texts: Case of Croatian," Springer, Studies in Computational Intelligence, Complex Networks V, vol. 549, 2014, pp. 275-283.

[26] D. Margan, A. Meštrović, and S. Martinčić-Ipšić, "Complex networks measures for differentiation between normal and shuffled Croatian texts," In Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, Croatia, IEEE, pp. 1819-1823, 2014.

[27] D. Margan, A. Meštrović. M. Ivašić-Kos, and S. Martinčić-Ipšić, "Toward a Complex Networks Approach on Text Type Classification," 6th International Conference on Information Technologies and Information Society, Slovenia, 2014.

[28] A. Masucci, and G. Rodgers, "Network properties of written human language," Physical Review E, vol. 74, no. 2, pp. 026102, 2006.

[29] A. Masucci and G. Rodgers. "Differences between normal and shuffled texts: structural properties of weighted networks". Advances in Complex Systems, vol. 12, no. 1, pp. 113-129, 2009.

[30] T. Milenković, and N. Pržulj, "Uncovering biological network function via graphlet degree signatures," Cancer Informatics, vol. 6, pp. 257273, 2008.

[31] M. E. J. Newman, Networks: An Introduction, Oxford University Press, 2010.

[32] W. De Nooy, A. Mrvar, and V. Batagelj, Exploratory social network analysis with Pajek. Cambridge University Press, Vol. 27, 2011.

[33] H. Rizvić, S. Martinčić-Ipšić, and A. Meštrović, "Network motifs analysis of Croatian literature," In Proceedings of the 6th International Conference on Information Technologies and Information Society, Slovenia, pp. 2-7, 2014.

[34] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," Genome Research, vol. 13, no. 11, pp. 2498-2504, 2003.

[35] R. V. Sole, B. Corominas-Murtra, S. Valverde and L. Steels, "Language networks: Their structure, function, and evolution," Complexity, vol. 15, no. 6, 2010, pp. 20-26.

[36] S. Šišović, S. Martinčić-Ipšić, and A. Meštrović, "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, pp. 1824-1829, 2014.

[37] S. Šišović, S. Martinčić-Ipšić, and A. Meštrović, "Toward network-based keyword extraction from multitopic web documents," In Proceedings of the 6th International Conference on Information Technologies and Information Society, Slovenia, pp. 18-27, 2014.

[38] S. Wernicke, and F. Rasche, "FANMOD: a tool for fast network motif detection," Bioinformatics, vol. 22, no. 9, pp. 1152-1153, 2006.

[39] Python Software Foundation. "Unicode Database Documentation," [Online]. Available: https://docs.python.org/2/library/unicodedata.html#unicodedata.normalize [Accessed: April 10, 2015].

[40] The Unicode Consortium. "Unicode Standard Annex #15: Unicode normalization forms," http://unicode.org/reports/tr15/#Norm_Forms [Accessed: April 10, 2015].