Two-fold complex network approach to discover the impact of word-order in Urdu language

Nuzhat Khan[1], Mohamad Anuar Kamaruddin[2], Usman Ullah Sheikh[3], Muhammad Paend Bakht[4]
[1]School of Industrial Technology, Universiti Sains Malaysia, Malaysia
[2,4]School of Electrical Engineering, Universiti Teknologi Malaysia, Malaysia
[1,3,4]Faculty of Information and Communication Technology, Balochistan University of Information, Technology, Engineering and Managaement Sciences, Pakistan

Article Info

Article history:

Received Mar 3, 2021
Revised Jun 29, 2021
Accepted Jul 6, 2021

Keywords:

Complex network
Small world
Urdu graph
Urdu word order
Word co-occurrence

ABSTRACT

This work examines standard Urdu text to confirm impact of word order in the language structure. The complex network approach is used to obtain universal properties of two different word co-occurrence networks. Macro and micro scale two-fold examinations of networks are performed for structure discovery. While preserving the vocabulary size, two networks are generated from same text with and without standard word order. In addition, text networks are benchmarked with a random network to extract global features. Achieved outcomes indicate certain word order in Urdu structure for most of the sentences. The normal and shuffled text networks demonstrated similar large-scale characteristics. The results show that average path length and network diameter is reduced after shuffling. On the other hand, clustering coefficient is increased in shuffled text as compared to normal text. Our results validated that few short sentences in range of three words are fully free order. The observations revealed that long sentences are ambiguous without standard order. Both networks are topologically similar but shuffling caused massive discrepancy in network composition and sentence structure. Inside graph view, grammatical association-based words connectivity exists in normal text network. With this universal approach, impact of word order in Urdu language is confirmed. Meanwhile, this breakthrough directs to uncover language composition by extracting small sentences as motifs.

Corresponding Author:

Mohamad Anuar Kamaruddin
School of Industrial Technology
Universiti Sains Malaysia
11800, Pulau Pinang, Malaysia
Email: anuarkamaruddin@usm.my

1. INTRODUCTION

The pioneer linguistic researchers have defined human language as a system of inter-reliant terms. It is composed of multiple interconnected linguistic units, commonly known as words. Every word falls out uniquely from the synchronized existence of the other words [1]. After deep analysis of this statement, linguistic research found a new direction towards revolutionary integration of complex networks. This framework has been proven a key approach in transformation of human languages into network for structure exploration. Resulting networks are appropriately deployed in advance machine learning and deep learning models [2]. Lexical networks bring together correlated components in the complex system of human language. These components include characters, words, and sentences depending on the nature of relationships in semantic networks, syntactic dependency networks or co-occurrence networks. Complex networks are some state-of-the-art models which exhibit small world property and scale free behaviour [3]. Some researchers confirmed small world and scale free properties by examining syntactic dependency between linguistic units and words co-occurrence networks construction [4]. Every model of lexical network displays complex network properties regardless of design, type and building techniques [5]. The above reviewed works have mainly focused on the twofold language network models on macro scale. These large-scale language structure investigations have successfully discovered the average path length, average clustering coefficient and betweenness centrality in a linguistic network [6]. Even without taking in to account fundamental grammatical rules and syntax of the languages, the complex network still reflects basic structure of the natural languages [7].

It has been verified through comparison of normalized and shuffled text networks that shuffled text can be distinguished from real text with help of some micro features presented in apparently macro scale models. Since human language is a highly complex system with some hidden patterns that can be explored via universal network techniques to model its complexity in a reliable, predictable, and measurable way [8]. Moreover, the massive application of micro, meso and macro scale patterns in a language structure can support machine for successfully producing correct sentences [9]. Besides the complex nature and a huge number of languages, the process of artificial language generation, content evaluation and translation are remarkable in terms of efficiency. The speed and accuracy of linguistic tasks are effect of small world structure to some extent. If properly discovered, these detectable patterns in linguistic structures will play an important role in language processing, creation, survival and existence as well. However, most of the previous research concentrated on English language processing, which resulted in evolution of modern English as a highly adopted and rich resource language. On the other hand, some Asian languages remained unexplored due to lack of required lexical resources and difficulty in processing. Consequently, most of current software and tools are not corresponding to these languages. On the other hand, Urdu is a highly spoken language around the world and regarded as national language of Pakistan. However, due to lack of sufficient research, this beautiful language is striving for its existence in modern internet era. Previous work has declared Urdu as a ‘low resource’ language [10]. Unavailability of required resources is the main reason of insufficient structure information of Urdu in current literature. As a result, most of language processing tools are incompatible with Urdu text structure and script. We found a dead lock between language processing tools and linguistic researchers during problem analysis and literature review [10]. It is observed that most of modern programming languages and tools do not properly support Unicode encoding UTF-8 [11]. This encoding system contains characters, numbers, and symbols in Persian-Arabic script of Urdu also known as Nastalik script [12]. For designing a language model, developers need sufficient linguistic resources and fine-grained relevant structural information to achieve the language compatibility with natural language processing (NLP) tools. Discovered deadlock cycle that lasted between linguistic researchers and NLP tools developers is clarified in Figure 1.

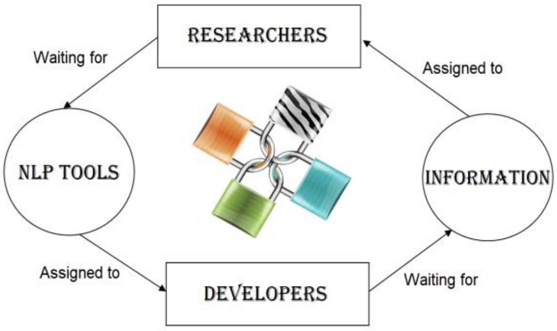


Figure 1. Deadlock between researchers and developers

This work aims to break the current cycle in order to open up different routes for innovative research on low resource languages. The analysis is conducted while keeping in mind above facts and position of low resource, poor languages that are facing some challenges in machine learning. Large amount of well-structured and pre-processed training set is a fundamental requirement of current machine learning models. Additionally, reliable labelling, significant informative features and powerful framework for application of modern techniques are also necessary. To bridge the research gap between Urdu language and contemporary machine learning, we have processed Urdu text for its structure discovery. This commenced work will lead to advanced study of Urdu structure within framework of complex networks theory for machine learning, deep leaning, and high-tech multilingual tools development. Moreover, there is a demand of structure discovery generally in the field of artificial intelligence and specifically, in the area of Urdu language processing. This huge gap can be visualized in Figure 2, which demonstrates the requirement of structure discovery and big data visualization in machine learning for natural language processing (NLP), multilinguistics, and artificial intelligence (AI).

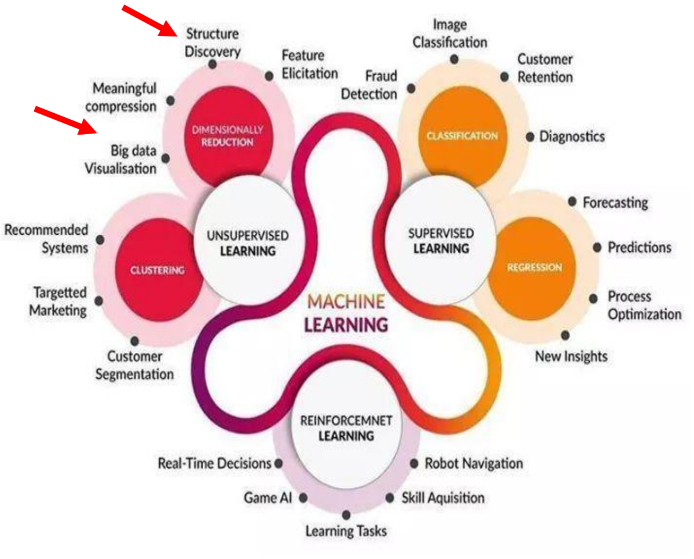


Figure 2. Requirement of structure discovery in machine learning

Structure discovery and big data visualization are basic building blocks to construct machine-learning models of real-world systems. In early 90's some relevant work for analysis of random networks was introduced in [13] while frequency distribution of words in a random text was investigated by [14]. The studies showed frequency distribution of words relevant to Zipfian law's statement. Therefore, it was concluded that in a natural language network, syntactic structure and scale freeness have no interdependence. In [15], it was also disovered those two networks of same size generated without specific order (created in random way) contain similar properties. Subsequently, another study examined the words frequency and sentence size relation in a co-occurrence network through shuffling [16]. This analysis was based on sentences taken as nodes and links between collocated nodes. Comparison of two random networks generated with syntactic and non-syntactic was investigated to measure the complexity of both networks in [17]. Accordingly, complex network structure of random text and linguistically organized text was investigated that determines the small world and scale free features in both networks [18]. Some works discussed the impact of word order, power law and degree distribution in a weighted network [7], [19], [20]. The aforementioned studies proved that in a weighted network of well-structured text, power of high frequency nodes remains equal to nodes in the shuffled text [21]. Although most of world languages follow universal features, still their differences are worth detecting. Our previous works explored complex network properties of Urdu [22] while two empirical linguistic laws are verified to look into statistical attributes of the language [23], [24]. The resulting features motivated to broaden investigation of word order in Urdu structure.

When it comes to overcome the difficulty in structural information extraction, networks have already been contributed with minimum preconditions. In this regard, graph theory is applied in different research fields to uncover the structure of systems containing connected components. It has strong association with so called 'network analysis' that played an important role in advancement of existing computer age. With increased interest in real world complex systems that are considered as networks, implementation of graph theory on such systems is also expanding [25]. Various kinds of networks were analysed for complex network properties in different fields. After deep analysis, such networks were demonstrating a number of giant components, a certain topological structure, and specific compositional patterns besides all core complexity [26]. Some successfully utilized complex networks include biological networks, technology networks, roads networks, social networks and most prevalent, language networks [27].

A lot of designing techniques are applied to construct lexical networks. But in case of exploring low resource language, it is technically preferable to pick co-occurrence network [28]. In fact, this network operates on co-existence of words, which is the only well-known relation among components of a plain text [29], [30].

2. FUNDAMENTALS OF NETWORK STRUCTURE

Any system comprising of connected components is established in the form of a network. Pictorial representation of real-world network turns into visual graph of the system. Visualization of resulting graphs gives deep insight into hidden patterns of system's components. Belonging to the field of graph theory, the network structure delivers some basic information about the core systems. Here are few comprehensive concepts related to network theory that need to explain ahead of its application. A simple graph *G* is expressed as (1).

$$G = (V, E) \tag{1}$$

The graph *G* consists of two finite sets *V* and *E* containing vertices and edges, respectively. The vertices are also called nodes while the connecting edges are commonly known as links. Every edge $e \in E$ is adjacent to two members 'v' and 'u' of set *V*. According to [31], in an undirected graph,

$$e = (v, u) where\ (v, u) = (u, v) \tag{2}$$

2.1. Graph geodesic

Number of direct edges between two nodes is considered as shortest path from source node to sink node [32]. Distance D between *i* nodes and *j* edges is symbolized as '*Dij*' which is in fact shortest path length from node *i* to node *j*. It is also called geodesic distance. One or more shortest paths may exist between two vertices [33].

2.2. Network density

In a network, portion of potential connections in actual connections is known as network density. It is formulated in [34], [35] as (3).

$$Network\ Density = \frac{Actual\ connections}{Potential\ connections} \tag{3}$$

2.3. Network size

A network *G* of density *d* with *N* nodes and *L* links emits size *N* that is equal to total number of nodes. This means network size is total number of nodes in the network [36].

2.4. Network window

This threshold value defines number of following co-occurred nodes that are permitted to connect. In case of co-occurrence network with predefined window, its window size is considered as maximum number of subsequent components, which are linked together to form the network [37].

2.5. Clustering coefficient

This is measure of nodes tendency in a graph to be clustered together. It can be calculated locally as well as globally corresponding to requirements and sort of analysis. In an undirected graph, clustering coefficient is equivalent to total number of triangles in the graph. Clustering coefficient in a network can be calculated with respect to individual node as local clustering coefficient. Average of all local clustering coefficients is taken as global clustering coefficient. Local and global clustering coefficients are calculated differently depending on type of the network [38]. The average clustering coefficient of a graph is measured with all local clustering coefficient of individual nodes. Clustering coefficient for 'i' node in an undirected graph is measured as (4).

$$C_i = \frac{number\ of\ triagles\ linked\ with\ mode\ \mathrm{i}}{number\ of\ triagles\ connected\ to\ immediate\ neighbors\ of\ mode\ \mathrm{i}} \tag{4}$$

The clustering coefficient for the entire graph is measured as average value of all the local clustering coefficients as (5).

$$GC = \sum_{i=1}^{N} C_i \tag{5}$$

GC represents global clustering coefficient while Ci is local clustering coefficient and N is number of nodes (network size). Value for clustering coefficient remains between 0 to 1. It is described in [39] according to (6) and (7).

$$0 \leq Ci \leq 1 \quad (6)$$

$$0 \leq GC \leq 1 \quad (7)$$

3. RESEARCH METHOD

For further proceeding to application of deep learning techniques on Urdu, it is obligatory to discover structural properties of this language first, without getting preoccupied by low resources. Staying on this fact, a model of co-occurrence text network is implemented on Urdu text corpus. An open corpus was carefully created from local newspapers, poetry books, novels and two religious books Quran and Bible translated in Urdu from Arabic and Hebrew, respectively. For this case, open corpus refers to generic text collection instead of discipline specific closed corpus. Details of initially collected machine-readable Urdu text are given in Figure 3.

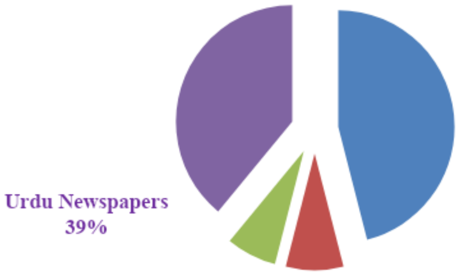
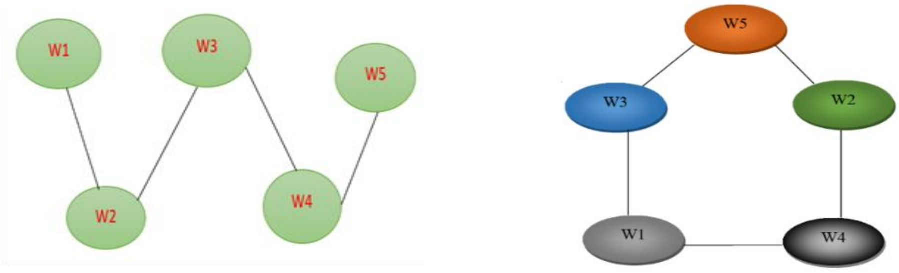


Figure 3. Corpus collection

An unannotated corpus is constructed after proper preprocessing and text cleaning [23]. Instead of annotation, each word was considered an independent entity regardless of grammer and meaning. Based on bag of bigrams model a word co-occurrence network is designed through sliding window technique [22]. Resultant network is converted to a graph *G* having more than 5K words as nodes. Edges are created as path/link between every two coexisting words in real word order. Another graph of same size is created from shuffled text to compare both networks in Python3.4 using networkx and matplotlib plotting libraries. Then, both network graphs are exported to Gephi 0.9.2 for visualization, analysis, and comparison.

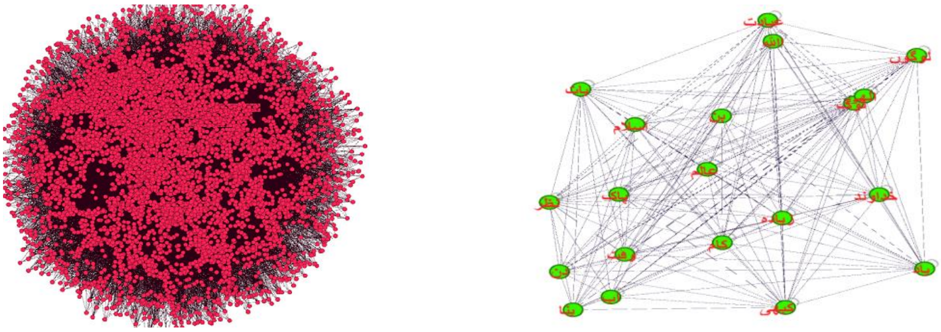
During text cleaning for corpus construction, multiple challenging exceptional features of Urdu script such as right to left text direction, multiple word forms, context sensitivity, incorrect spacing, undefined word boundary and spelling mistakes were handled using term frequency-inverse document frequency (TF-IDF) technique as adopted in [23]. Previously, Urdu was stated a free order language earlier, which does not follow certain grammatical rules and particular word order [10]. This misinterpretation was because of not performing scientific research for exploration of word order and sentence structure in natural Urdu language [40]. In our previous work [22], it was observed that there are some hidden patterns in complex topological structure of Urdu network expressed in the form of word order. Specific stable order was detected in most of long sentences except for some short sentences. To confirm the existence of well-preserved word order, we transformed Urdu text into two different co-occurrence networks for its graphical exploration. One network is constructed by connecting co-occurred words according to order provided in the original text. For example, five words generating a sentence as W1 W2 W3 W4 W5, where W denotes specific words and digits indicate position /index of each word in the sentence. The second network is constructed from shuffled text by altering original word order randomly. The words association methods in normal and shuffled network are shown in Figure 4(a) and (b), respectively.

In both networks, the same text sample is utilized for evaluation. The size of vocabulary is preserved and number of nodes in both networks are equal. Number of nodes are more than 5K in normalized and shuffled text network. To confirm small world and scale free property, both networks are compared with Erdos-Renyi random graph. Text networks are generated likewise in terms of production method however, the text for shuffled network was randomly rearranged. The co-occurrence network graph in default view is shown in Figure 5(a) while zoomed view of labeled nodes is displayed in Figure 5(b).



(a) (b)

Figure 4. Toy networks; (a) normalized text and (b) shuffled text



(a) (b)

Figure 5. Co-occurrence network; (a) default layout and (b) zoomed view of labelled nodes

4. RESULTS AND DISCUSSION

Original order of text remained preserved in network of normal text but in the shuffled text network, all nodes were readjusted in random way. Although new shuffled network lost typical natural language word order, but topological structure still exhibits small world behaviour as shown in Table 1. Apparently reordered text network attained lowered diameter as compared to normalized standard network. On the other hand, diameter of shuffled network equates the random network. Reduced number of links, diminished communities and decreased size of network diameter could be due to the loss of relevancy among co-occurred words. Anticipated higher clustering coefficient in shuffled network as compared to normal text differentiates the original text from the one without preserved order. Macro scale analysis of both networks on basis of universal topological features indicated a strong effect of shuffling through the altered structure. The difference between both networks with some underlying structural parameters clarifies that 'definitely some specific word-order exits in Urdu language'. Although order freeness is spotted for few very short sentences only, which means long sentences cannot convey accurate connotation/message without a particular word-order. For clarification of the outcome, two sample sentences of different size are arranged in figure short sentence composed of three words given in Figure 6(a) appears to be almost free of order. Any combination of three words accumulated in the sentence seems to covey the similar message. This is a good feature to be used in machine leaning but unfortunately, long sentences do not follow this order-free arrangement. A bit longer sentence containing five words is considered correct if and only if words are composed in anti clock order starting from upper most word as illustrated in Figure 6(b). Although some other word-order combinations can produce correct sentence, but appropriate standard composition needs to be regarded in this case.

Table 1. Global features of networks

Properties	Normal graph	Shuffled graph	Random graph
No of nodes	5180	5180	5180
No of Links	101415	101397	101415
Diameter	5	4	4
Radius	3	3	3
Average path length	2.63707629	2.6370723	2.73062725
Clustering coefficient	0.386	0.399	0.007
Number of communities	11	8	1

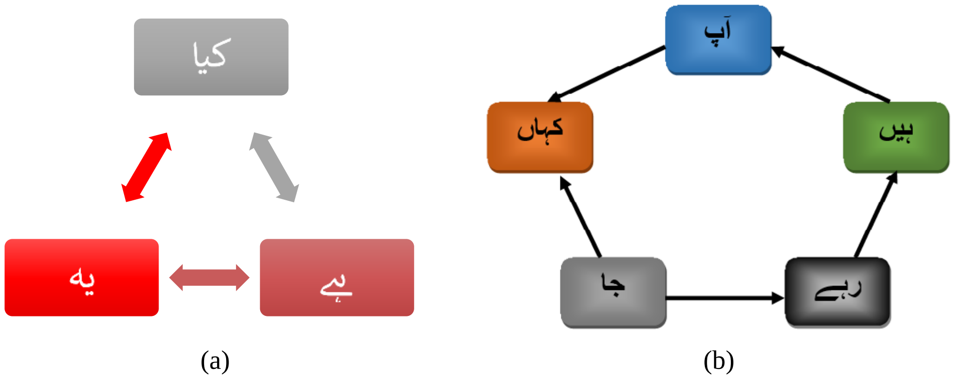


Figure 6. Word order; (a) short sentence and (b) long sentence

Although the global features and examined sentence structure are distinguishing between ordered and shuffled networks. By putting some extra effort on both networks for deeply examining individual nodes and links inside networks, few more interesting patterns are detected. As different hubs (centred nodes) in both networks are filtered along with immediate neigbours, subgraphs in both networks display major difference. This process helped in examining network's structure and nodes behaviour on micro level. Not only immediate neighbours are diverse in both networks in terms of connectivity, but shuffling has also wiped out the grammatical association among nodes. In shuffled network, linked nodes neither present any grammatical association with centerd word (hub) nor display any relevance to each other. While nodes are grammatically attached to hub as well as topically correlated with other connected words in normal text netwok. A little insight into both networks revealed that words in shuffled network were linked without grammatical association. Due to lack of association among connected words, nodes association remained random and unobvious. However, the nodes in normal text network were revealing some grammatical relationships among most of adjacent links. Not all but somehow, most of direct neighbors to centred node were topically relevant to hub as well. On deep analysis of centered nodes (hubs), it is observed that the immediate neigbours not only differ in both networks but also reveal consistent behaviour subject to the network. To further clarify the observed patterns, we illustrate it by focusing on unique targeted node (خدا) which is a word from original Urdu text. The centred node is Urdu word for 'God' and it is linked with words 'obidience', 'repent', 'creature', 'reward', 'believers', 'mighty', 'grateful', 'blessings' and 'creator'. Every node when analysed makes sense that it is directly connected to its co-occurred words from same context. The immediate neighbours are not only associated on the bases of coexistence in the text, but labels of nodes reflect their grammatical connection, context cantered organization and topic relevancy with the hub as well as with each other. This association is lost in shuffled network as indicated in Figure 7(a) that the same word is connected with fewer nodes. The word 'God' has only three immediate neibour nodes with lable 'wide', 'sentences' and 'throat' which appear to be less interrelated on basis of context. The centered node and immidaite neighbors in both text networks are presented in Figure 7(a) and (b).

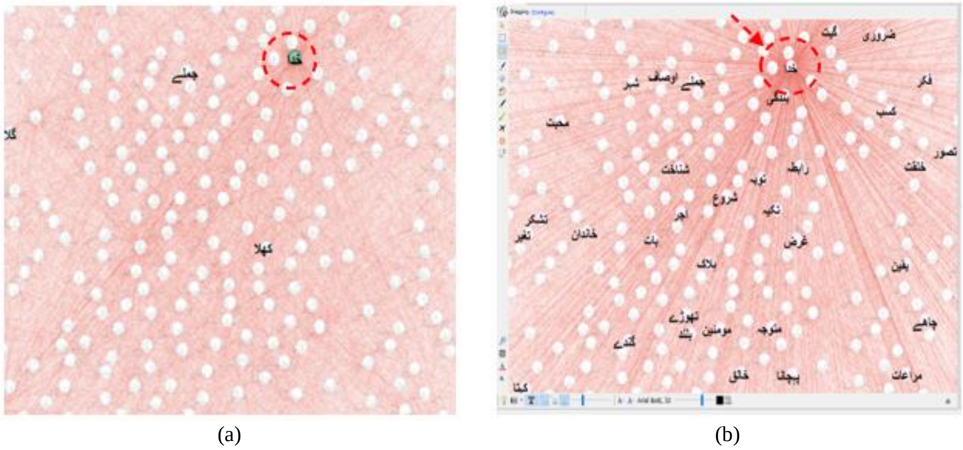


Figure 7. Centered word with associated nodes; (a) shuffled text and (b) normal text

Mainly three basic structural differences are found from micro scale analysis of isolated subnetwork in both graphs. The analysed nodes are different in terms of node degree, centrality, and topic relationship among linked nodes as well as with centred nodes. The studied nodes are hubs in normal text due to high degree while in shuffled text, number of links to the nodes have been significantly reduced. Consequently, nodes degree distribution patterns evolved from hierarchical to marginally randomized manner after shuffling. These phenomena also confirm the existance of word order in topological and grammatical structure of Urdu language.

5. CONCLUSION

In this work, two co-occurrence networks of normal text and shuffled text are constructed from Urdu corpus written in Nastalik script. The corpus is collected from four machine-readable text sources; Urdu books, Newspapers, Holy Quran, and Holy Bible. Bag of bigram model is developed for words networks construction. Afterwards, the text networks composed of more than five thousand words each are transformed to graph. With graph visualization and statistical analysis, impact of word order is examined to verify that Urdu is not a free order language. It is found that global features of both networks exhibit small world and scale free properties. However, shuffeling reformed the words connection patterns. The results based on empirical data and graph visualization confirmed that Urdu holds solid word order. After deeply investigating both networks, it is analyzed that shuffling evaporated grammatical associations among connected words. The study suggests that Urdu cannot be treated as a fully free order language. In this regard, Urdu needs to be investigated for understandable structural patterns in its word order. This study can be expanded by applying state-of-the-art structure mining techniques on Urdu text. For future research, we recommend motifs extraction algorithm involving motifs frequency to capture appropriate word order in Urdu sentence structure including stop words.

REFERENCES

[1] W. P Goh, K.-K. Luke, and S.A. Cheong, "Functional shortcuts in language co-occurrence networks," *PLoS ONE*, vol. 13, no. 9, pp. e0203025, 2018, doi: 10.1371/journal.pone.0203025.

[2] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824-827, 2002, doi: 10.1126/science.298.5594.824.

[3] R. F. I. Cancho and R.V. Solé, "The small world of human language," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 268, no. 1482, 2001, pp. 2261-2265, doi: 10.1098/rspb.2001.1800.

[4] T. Linzen and B. Leonard, "Distinct patterns of syntactic agreement errors in recurrent networks and humans," *arXiv preprint arXiv:1807.06882*, 2018.

[5] T. Stanisz, J. Kwapień, and S. Drożdż, "Linguistic data mining with complex networks: a stylometric-oriented approach," *Information Sciences*, vol. 482, pp. 301-320, May 2019, doi: 10.1016/j.ins.2019.01.040.

[6] A. Criado-Alonso, E. Battaner-Moro, D. Aleja, M. Romance, and R. Criado, "Using complex networks to identify patterns in specialty mathematical language: a new approach," *Social Network Analysis and Mining*, vol. 10, no. 1, pp. 1-10, 2020, doi: 10.1007/s13278-020-00684-1.

[7] A. P. Masucci and G. J. Rodgers, "Differences between normal and shuffled texts: structural properties of weighted networks," *Advances in Complex Systems*, vol. 12, no. 01, pp. 113-129, 2009, doi: 10.1142/S0219525909002039.

[8] B. Corominas-Murtra, R. Hanel, and S. Thurner, "Understanding scaling through history-dependent processes with collapsing sample space," *Proceedings of the National Academy of Sciences*, vol. 112, no. 17, pp. 5348-5353, 2015, doi 10.1073/pnas.1420946112.

[9] J. Wang, "Generating An Overview Report of Multilevel Structure over A Large Corpus of Documents," Doctor Dissertation, Department of Computer Science, University of Massachusetts Lowell, USA, 2019.

[10] A. Daud, W. Khan, and D. Che, "Urdu language processing: a survey," *Artificial Intelligence Review*, vol. 47, no. 3, pp. 279-311, 2017, doi: 10.1007/s10462-016-9482-x.

[11] S. Hussain, N. Durrani, and S. Gul, "Survey of language computing in Asia," *Center for Research in Urdu Language Processing*, National University of Computer and Emerging Sciences, vol. 2, pp. 2005, 2005.

[12] S. Shabbir and I. Siddiqi, "Optical character recognition system for Urdu words in Nastaliq font," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 5, pp. 567-576, 2016, 10.14569/IJACSA.2016.070575.

[13] J. Sirosh and R. Miikkulainen, " Self-Organization and Functional Role of Lateral Connections and Multisize Receptive Fields in the Primary Visual Cortex," *Neural Processing Letters*, vol. 3, no. 1 pp. 39-48, 1996.

[14] W Li, "Random texts exhibit Zipf's-law-like word frequency distribution," *IEEE Transactions on information theory*, vol. 38, no. 6, pp. 1842-1845, Nov. 1992, doi: 10.1109/18.165464.

[15] R. Popping, *Computer-assisted text analysis*, Los Angeles, United States: Sage, 2000.

[16] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *Nature*, vol. 393, no. 6684, pp. 440-442, 1998, doi:10.1038/30918.

[17] H. H. Liu and F. Hu, "What role does syntax play in a language network?," *EPL (Europhysics Letters)*, vol. 83, no. 1, pp. 18002, 2008, doi: 10.1209/0295-5075/83/18002.

[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] A. P. Masucci and G. J. Rodgers, "Network properties of written human language," *Physical Review E*, vol. 74, no. 2, pp. 026102, 2006, doi: 10.1103/PhysRevE.74.026102.

[20] D. R Amancio, "Probing the topological properties of complex networks modeling short written texts," *PloS ONE*, vol. 10, no. 2, pp. e0118394, 2015, doi: 10.1371/journal.pone.0118394.

[21] D. Margan, A. Mestrovic, and S. Martinčić-Ipšić, "Complex networks measures for differentiation between normal and shuffled Croatian texts," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, IEEE, 2014.

[22] N. Khan, M. P. Bakht, M. J. Khan, and A. Samad, "Complex Network of Urdu Language," *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, 2019, pp. 1-6, doi: 10.1109/MACS48846.2019.9024791.

[23] N. Khan, M. P. Bakht, and R. A. Wagan, "Corpus Construction and Structure Study of Urdu Language using Empirical Laws," Urdu News Headline, Text Classification by Using Different Machine Learning Algorithms, 2019.

[24] N. Khan, M. P. Bakht, M. J. Khan, A. Samad, and G. Sahar, "Spotting Urdu Stop Words By Zipf's Statistical Approach," *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, 2019, pp. 1-5, doi: 10.1109/MACS48846.2019.9024817.

[25] 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," *2008 IEEE International Symposium on Parallel and Distributed Processing*, 2008, pp. 1-12, doi: 10.1109/IPDPS.2008.4536261.

[26] G. Chen, X. Wang, and X. Li, *Fundamentals of complex networks: models, structures and dynamics*, John Wiley & Sons, 2014.

[27] N. E. Eikmeier, "Spectral Properties and Generation of Realistic Networks," Thesis, Department Mathematics, Purdue University Graduate School, United States, 2019, doi: 10.25394/PGS.8340911.v1.

[28] D. Margan and A. Meštrović, "LaNCoA: A Python toolkit for Language Networks Construction and Analysis," *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015, pp. 1628-1633, doi: 10.1109/MIPRO.2015.7160532.

[29] H. Li, H. An, Y. Wang, J. Huang, and X. Gao, "Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: Based on two-mode affiliation network," *Physica A: Statistical Mechanics and its Applications*, vol. 450, pp. 657-669, 2016, doi: 10.1016/j.physa.2016.01.017.

[30] C. Akimushkin, D. R. Amancio, and O. N. Oliveira Jr, "Text authorship identified using the dynamics of word co-occurrence networks*," PloS ONE*, vol. 12, no. 1, pp. e0170527, 2017, 10.1371/journal.pone.0170527.

[31] A. Abboud, R. Krauthgamer, and O. Trabelsi, "New algorithms and lower bounds for all-pairs max-flow in undirected graphs. in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp.48-61, SIAM, 2020, doi: 10.1137/1.9781611975994.4.

[32] C. Daniel, A. Furno, and E. Zimeo, "Cluster-based Computation of Exact Betweenness Centrality in Large Undirected Graphs," *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 603-608, doi: 10.1109/BigData47090.2019.9006576.

[33] Y. Y. Adikusuma, Z. Fang, and Y. He, "Fast Construction of Discrete Geodesic Graphs," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 2, pp. 1-14, 2020, doi: 10.1145/3144567.

[34] K. Dong, A. R. Benson, and D. Bindel, "Network density of states," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019.

[35] B. S. Graham, F. Niu, and J. L. Powell, "Kernel density estimation for undirected dyadic data," *arXiv preprint arXiv:1907.13630*, 2019.

[36] D. Bucur and P. Holme, "Beyond ranking nodes: Predicting epidemic outbreak sizes by network centralities," *PLOS Computational Biology*, vol. 16, no. 7, pp. e1008052, 2020, doi: 10.1371/journal.pcbi.1008052.

[37] W. A. Alzoubi, "An Improved Graph based Rules Mining Technique from Text," *Engineering World Open Access Journal*, vol. 2, 2020.

[38] M. Savić, M. Ivanović, and L. C. Jain, "Fundamentals of Complex Network Analysis," in *Complex Networks in Software, Knowledge, and Social Systems*, *Springer*, pp. 17-56, 2019, doi: 10.1007/978-3-319-91196-0_2.

[39] L. Iskhakov, B. Kamiński, M. Mironov, P. Prałat, and L. Prokhorenkova, "Local clustering coefficient of spatial preferential attachment model," *J. of Complex Nets.*, vol. 8, no. 1, pp. cnz019, 2020, doi: 10.1093/comnet/cnz019

[40] N. S. Naz, K. Hayat, M. I. Razzak, M. W. Anwar, S. A. Madani, and S. U. Khan, "The optical character recognition of Urdu-like cursive scripts," *Pattern Recognition*, vol. 47, no. 3, pp. 1229-1248, 2014, doi: 10.1016/j.patcog.2013.09.037.

BIOGRAPHIES OF AUTHORS

Nuzhat Khan received her MS degree in Information Technology from Balochistan University of Information Technology, Engineering and Management Science (BUITEMS) Quetta Pakistan in 2019 through Higher Education Commission (HEC) scholarship. Currently, she is pursuing her PhD degree in Department of Environmental Technology Universiti Sains Malaysia. Her research interests include Natural Language Processing, Applied Linguistics, Statistical Linguistics, Network Science, Graph Theory, Artificial Intelligence, Machine Learning and Deep Learning. She is recently involved in research on application of Artificial Intelligence in oil palm industry.

Dr. Anuar received his BSc (Civil Engineering), MSc and PhD in Environmental Engineering from Universiti Sains Malaysia (2005-2015). His research interests include on waste management, landfill leachate treatment and industrial wastewater treatment. Dr. Anuar also has participated in several international conferences organized by reputable bodies worldwide. He also actively engaged in community empowerment through multidisciplinary research that allow him to communicate better with the bottom billion. Dr. Anuar has published peer review articles in various reputable publishers. Dr. Anuar also actively involved in numerous technical visits and environmental protection works related to his expertise. He also sits in various organizing committees, technical committee and a member of several prestigious bodies worldwide.

Dr. Usman Ullah Sheikh received his BEng degree (2003) in electrical and mechatronics engineering, the MEng degree (2005) in telecommunications engineering and PhD degree (2009) in image processing and computer vision from Universiti Teknologi Malaysia. His research work is mainly on computer vision and embedded systems design.

Muhammd Paend Bakht received his MS degree in Telecom Engineering from Balochistan University of Information Technology, Engineering and Management Science (BUITEMS) Quetta Pakistan in 2016 through Higher Education Commission (HEC) scholarship. Currently, he is pursuing his PhD degree in School of Electrical Engineering, Universiti Teknologi Malaysia. His research interests include optimization and energy management of grid connected renewable energy systems, smart grids and forecasting of solar and wind turbine power**.**