Journal of Complex Networks (2019) **7**, 114–127
doi: 10.1093/comnet/cny018
Advance Access Publication on 13 August 2018

Applying network theory to fables: complexity in Slovene belles-lettres for different age groups

RENE MARKOVIČ

Department of Physics, Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia, Faculty of Education, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia and Faculty of Energy Technology, University of Maribor, Hočevarjev trg 1, SI-8270 Krško, Slovenia

MARKO GOSAK

Department of Physics, Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia and Faculty of Medicine, University of Maribor, Taborska 8, SI-2000 Maribor, Slovenia

MATJAŽ PERC

Department of Physics, Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia, CAMTP—Center for Applied Mathematics and Theoretical Physics, University of Maribor, Mladinska 3, SI-2000 Maribor, Slovenia and Complexity Science Hub, Josefstädter straße 39, A-1080 Vienna, Austria

MARKO MARHL

Department of Physics, Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia, Faculty of Education, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia and Faculty of Medicine, University of Maribor, Taborska 8, SI-2000 Maribor, Slovenia

AND

VLADIMIR GRUBELNIK[†]

Department of Physics, Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia and Faculty of Electrical Engineering and Computer Science, University of Maribor, Koroška cesta 46, SI-2000 Maribor, Slovenia
[†]Corresponding author. Email: vlado.grubelnik@um.si

Edited by: Ernesto Estrada

[Received on 27 June 2018; editorial decision on 12 July 2018; accepted on 14 July 2018]

Words are the building blocks of human communication. They are arranged in sentences in a non-trivial and universal way, which implies the existence of fundamental organizational principles that have shaped language development. One of the fundamental examples is the Zipf's law, which says that the frequency of word occurrence is roughly an inverse power-law function of its rank. In our article, we study the structure and complexity of texts in Slovene belles-lettres, with an emphasis on evaluating the differences in the texts for different age groups. We show that the co-occurrence connectivity of words forms a complex and heterogeneous network that is characterized by an efficient transfer of information. Moreover, we show that with the increasing age of readers, the length of texts, the average length of words, different combinations of phrases and the complexity of social interactions between literary characters, also all increase. Conversely, the fraction of unique words decreases. We also show that the word co-occurrence networks for older age groups are less modular and exhibit a higher level of small-worldness, and that the Zipf's exponent as a measure for the linguistic complexity is more negative, than for younger age groups. Taken together, we demonstrate that network theory enables an in-depth theoretical exploration of Slovene belles-lettres, with clear distinctions in statistical properties between different age groups, thus bridging art and exact sciences in a mutually rewarding way.

Keywords: complex networks; human language; linguistic typology and classification; syntactic structure; text processing; Zipf's law.

1. Introduction

The expression power of human language is limitless and a key step was the transition from a non-syntactic to a syntactical form of communication [1]. Syntactic communication is based on a combination of discrete elements to generate information. In case of human communication, these discrete elements are words, each of which has its own specific meaning. By combining different words, we create messages [2, 3]. This way of communication in turn enables us theoretically to create endlessly many ways to combine words and generate information's [3]. In contrast to the syntactic language used by humans, non-syntactic communication predominates the animal world. This form of communication is based on signals, whereby each signal has its specific meaning and as such represents a message [4]. It is well known that our way of communication has evolved out of a non-syntactic form of communication [5]. It has been shown that a transition from a non-syntactic to a syntactic form of communication is an evolutionary drive towards a more efficient and flawless information exchange [3, 6] and has occurred due to the high number of relevant and required signals between individuals in human society [2]. Independently of the form of communication (syntactic or non-syntactic), it has been empirically inferred that the frequency of occurrence of building blocks of communication decays as a power-law function of its rank with an exponent close to -1, which is known as the Zipf's law [4, 7, 8]. Zipf's rule dictates that most words appear very rarely and only a few words appear very frequently. The question emerged, what mechanism could cause such a universal behaviour. Many different competing explanations have been proposed, from the later dismissed ideas of random processes [9–11], to preferential repetitions or proportional growth [12, 13] and the principle of least effort of information transmission between the speaker and listener [8, 14].

More decisive statistical approaches and the incorporation of large corpora have made the evaluation of complexities of the word frequency distribution more precise. Improvements and generalizations of the simple Zipf's rule have later been proposed to account for the variances from pure power-law behaviour [15, 16]. While the precise mechanism driving the features of word frequencies and the structures beyond the Zipf's law are a matter of ongoing research, the simple power-law description captures well the scaling properties and complexity of the language, and is therefore, one of the most commonly used measures in quantitative linguistics [17]. Moreover, while the occurrence of at least near-Zipfian behaviour is universal across languages, the Zipf's exponent values were found to be language-dependent [18–20]. The Zipf's exponent was also found to change during language ontogeny and its value are less negative in children's speech in comparison with adults [21, 22].

The Zipf's law provides global information about the statistical features of language and reveals its complex character. However, for a more in-depth investigation of its structure and interactions between words and the relation to cognition, more advanced quantitative tools are required. In the last years, network science has been proposed as one of the most promising and unifying frameworks for probing the complexity of language and putting the linguistic research in a new and interdisciplinary context [23–25]. In the seminal paper of Cancho and Solé [26] a network of words was constructed on the basis of co-occurrences in sentences and was found to manifest exceptional topological features such as the small-world effect and a scale-free structure. Those findings were later confirmed and expanded for different languages and network construction techniques [27–31]. Besides syntactic and co-occurrence based networks, also conceptual and semantic networks of words were constructed with the aim to provide a more cognitive-oriented point of view [25, 32–34]. Notably, those networks were also found to express the previously mentioned complex topological features, such as power-law behaviour and a very small average shortest path lengths [33, 35, 36], although they were found to differ from syntactic networks in their hierarchical properties [37]. Furthermore, in a more content-inspired aspect, the network analysis proved to be a powerful tool in capturing interaction structures and dynamics of a story. By this means 'social network' interactions are extracted from literary texts based on co-occurrences and the resulting character networks can then be further analysed by conventional network approaches. It has been found that fictional social networks are small-world, highly clustered, and hierarchical, as real social networks [38, 39], but typically differ from real ones in connectivity and levels of assortativity [40]. Most importantly, these empirically network-based analyses of character interactions have been proven to successfully find communities, identifying influential characters, and evaluating the complexity of a storyline [41–43].

Till today, quantitative and network-based analyses on how the characteristics of literary texts change regarding the recommended age of readers, have not yet been performed. We address this issue in the present article by systematically analysing 53 Slovene belles-lettres for children and juveniles categorized into four different age groups. In each subgroup, we study the statistical properties of the language, from the basic to the more advanced measures. The later encompass the investigation of word co-occurrence structures and the resulting complex network analyses. Finally, we additionally construct and evaluate character networks to assess the complexity of interaction structures in stories in texts for different age groups.

2. Materials and methods

In this section, we describe the methodology to assess the statistical peculiarities of texts. The initial quantifications encompassing classical word appearance statistics and Zipf's law are followed by description of language networks generation and analysis with the aim to capture fundamental traits of language under a global picture. Finally, we focus on the construction of character networks that are used to capture the interaction structures in different fables.

2.1 *Dataset*

The original list of belles-lettres was created on the basis of recommended books for home reading, which were composed from experts in the field of linguistics at 16 elementary schools and kindergartens. From the set of 128 works on the list, we selected 53 on the basis of availability. Table 1 features the details. The combined texts contained 1 020 805 words (with repetition). We have compared the words from the 53 selected texts with the words listen in the Slovenian dictionary. Our analysis showed that 87.6% of the words recorded in the dictionary of the Slovenian literary language dictionary also appeared in the extracted word list.

Table 1 *The categorization of 53 texts used in our study*

Age group	Year range	Number of books
AG1	1–5	11
AG2	6–8	15
AG3	9–11	13
AG4	12–14	14

2.2 *Zipf's exponent*

The frequency of word occurrence in a text has been shown to follow a power-law, which is known as the Zipf's law. The law states that by sorting words based on the frequency of usage (i.e. most frequent word is assigned rank 1, the second most frequent word is assigned rank 2, ...), the resulting distribution obeys:

$$N(r) \propto r^{-\gamma}, \tag{1}$$

where $N(r)$ is the frequency of a word with rank r and γ is the exponent of the law also known as the Zipf's exponent. Herein, we study variations of γ in different texts that are recommended for readers in a certain age group. For this purpose, texts in a given age group are merged into a single text to extend the corpus of words and compute the exponent on a larger population of words. For the merged text, we then assign ranks to the words and lastly fit a power-law function to the word frequency distribution to assess the value of γ.

2.3 *Construction of language networks*

First, we import a given text file and split the text into individual sentences. Sentences are separated based on punctuation marks, which are used to mark the end of a sentence (i.e. "!","?","."). Then, we construct word co-occurrence networks in which individual words represent nodes and edges between them are established between two neighbouring words within a sentence, similarly as described previously [27–31, 44]. By this means a $N \times N$ adjacency matrix, $\mathbf{A}$, is constructed, where N is the number of unique words. Schematically, this process is illustrated in Fig. 1 with a sample text, which is merely used to demonstrate how the algorithm works.

We compute the length of each text, i.e. number of all words, N_L and the number of unique words N_{UW}. The ratio between those two parameters defines the unique word density η:

$$\eta = \frac{N_{UW}}{N_L}. \tag{2}$$

The values of parameter η span between $1/N_L$ and 1, whereby 1 corresponds to a text, where every word is a unique word ($N_{UW} = N_L$). On the other hand, smaller values of η indicate that words are frequently reused in the text. We also analyse the word length distribution, whereby the length of a word, l, is defined by the number of characters it is made of. In addition, we will also examine how frequently two neighbouring words with a given length appear within a text.

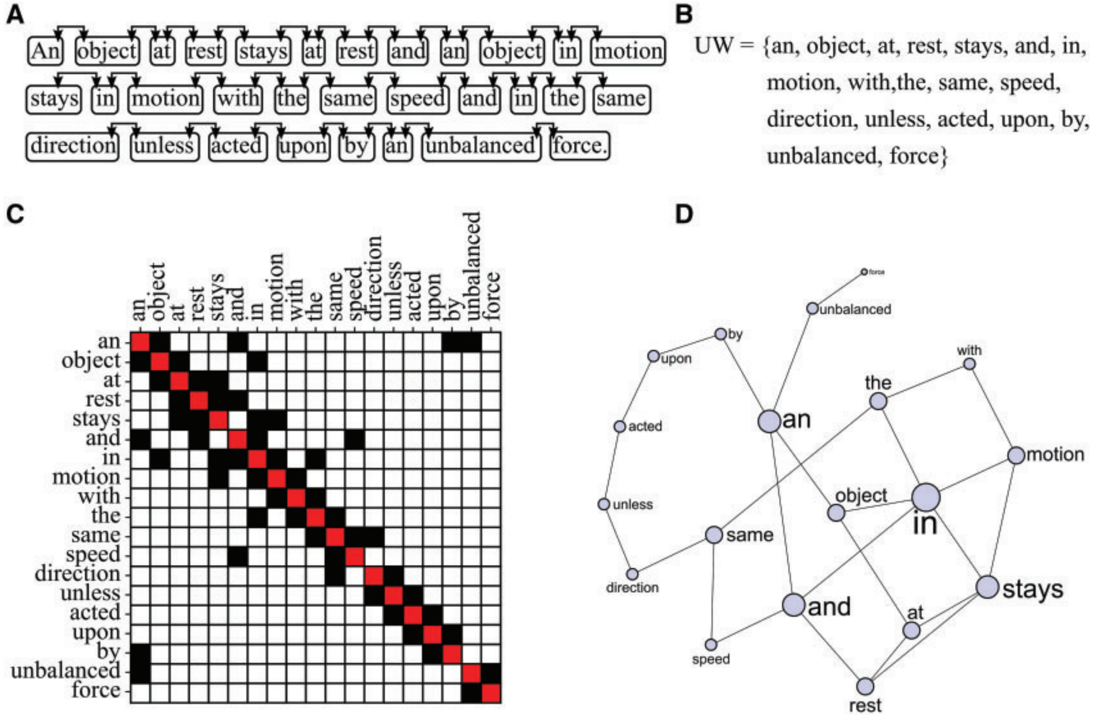
A
An object at rest stays at rest and an object in motion
stays in motion with the same speed and in the same
direction unless acted upon by an unbalanced force.
B
UW = {an, object, at, rest, stays, and, in, motion, with,the, same, speed, direction, unless, acted, upon, by, unbalanced, force}
C
an
object
at
rest
stays
and
in
motion
with
the
same
speed
direction
unless
acted
upon
by
unbalanced
force
D
force
unbalanced
by
upon
acted
unless
direction
speed
same
an
the
with
motion
object
in
stays
and
at
rest

FIG. 1. Language network creation. Sample text (A), the corresponding unique word list (B), adjacency matrix for the sample text (C) and the extracted language network for the sample text (D).

2.4 *Character networks*

To evaluate the complexity of interactions among the characters in a given text, we first create a list of all the main characters. Afterwards our script reads the text sentence by sentence and each sentence is indexed by integers. By this means, we define the ‘location’ of a given sentence within the text. Within a given sentence, we search if any of the characters appears in it. If a character is found, we store the corresponding sentence number. After this process, we have a sorted list with the information in which sentences the characters have been found. The list is then used to compute the distances among characters. We define this distance as the absolute difference between the appearances of any two characters. However, the frequency of character occurrence is related to the importance of character and hence the number occurrences of a given character name varies. Consequently, before computing the distances between two characters, we define the smallest number of occurrences, which also defines how many times, the two characters could be treated as connected. Afterward, we compute the corresponding distances, whereby the smallest number is taken as the distance between the two characters. For example, character X appears in sentences 1, 10, 20, 51, 56 and character Y appeared in sentences 15, 17, 31. We take the character Y, which is less frequently represented in the text, and calculate the minimum distances of the character X to 15, 17 and 31. The computed distances between the characters are 5, 3, 11. Hence, three is taken as the distance between them. The final character network, representing interactions among the main characters (sociogram), is then visualized and quantified by several network metrics, which are in more detail explained in the next subsection.

2.5 *Network metrics*

To characterize the topological features of the constructed networks, we will compute several network metrics, which will give a broad insight into the connectivity patterns of a given graph. The degree of a node is the most fundamental feature. It equals the number of edges the node has in the network. The degree of the i-th node, k_i, is defined as:

$$k_i = \sum_{j}^{N} A_{ij}. \tag{3}$$

It should be noted that many real-world networks display a scale-free degree distribution, with power-law exponents ranging from 2 to 3 [45]. Another metric, which we will compute for the networks is the average clustering coefficient C. This feature characterizes the presence of highly interconnected groups of neighbouring nodes. Those highly interconnected nodes are characterized with a high local clustering coefficient C_i. We implemented the method introduced by Watts and Strogatz [46], to compute the local clustering coefficients C_i. The local clustering coefficient C_i of the i-th node is defined as:

$$C_i = \frac{2n_i}{k_i(k_i - 1)}. \tag{4}$$

The local clustering coefficient is given as the ratio between the number of existing edges among the neighbours of the i-th node, n_i and the number of all possible edges among them, $k_i(k_i - 1)/2$. The average clustering coefficient is then defined as the average of all local clustering coefficients. Since the clustering coefficient quantifies the degree of a network to form local groups of highly connected nodes, it is also an indicator of the degree of segregation in a network. In contrast to the clustering coefficient, the average shortest path length, L, characterizes the network's integration of individual nodes [46]. To compute L, one must compute the length of all the shortest paths L_{ij} between all pairs of nodes in the network. Afterwards, L is computed as follows:

$$L = \frac{1}{N\,(N-1)} \sum_{i \neq j}^{N} L_{ij}. \tag{5}$$

From the shortest path length between all the nodes we can also compute the diameter, D, of the network, which is equal to the longest of all the calculated shortest paths in a network. Lastly, we will compute the modularity, Q, of a graph [47]. This is a commonly used measure to quantify how successful the partitioning of a graph into communities was and is defined as:

$$Q = \frac{1}{2m} \sum_{i,j}^{N} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{6}$$

where m is the total number of edges in the network, c_i is the membership variable, which assigns the i-th node to a community c_i and $\delta(c_i, c_j)$ is the delta function, which equals 1 if $c_i = c_j$ and 0 otherwise. The aim of the algorithm is to maximize the modularity by continuously reshaping the community structure of the network and replacing nodes among the communities. A new configuration is accepted if the gain

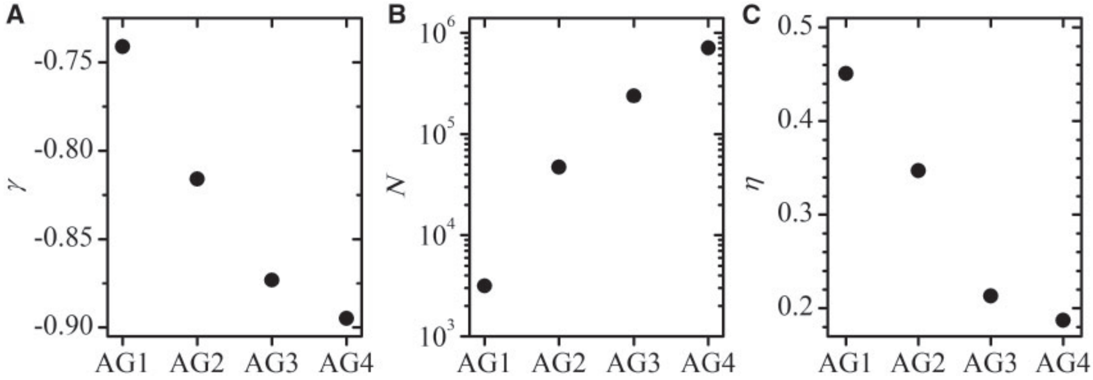

FIG. 2. Average features of texts in a given age group. Zipf's exponent (A), number of words (B) and unique word density (C).

in modularity ΔQ is positive. The process is repeated until no further improvement in ΔQ is achieved and the most likely community structure of the network is found.

Furthermore, small-world network features have been found in a variety of real world networks. The main topological features of such networks are a sparse connectivity, cliquishness and a short average path length. To quantify the degree of small-worldness, we numerically compute the small-word parameter S [48]. This parameter evaluates the ratio between the normalized network clustering and the normalized shortest path length, with respect to the equivalent random network. The higher the value of S is the more 'small-world' the network is. The parameter S is computed as:

$$S = \frac{C/C_{\text{rand}}}{L/L_{\text{rand}}}, \tag{7}$$

where C and L are the clustering coefficient and the average shortest path length of a given network, respectively, whereas C_{rand} and L_{rand} stand for the clustering coefficient and the average shortest path length of the corresponding random network with the same average degree and number of nodes.

3. Results

In our study, we analysed statistical features of 53 literary works categorized into different age groups (AGs). Details about individual AGs is given in Table 1. The combined texts where made out of 1 020 805 words (with repetition). By comparing the words with the Slovenian dictionary, we found 87.6% of the words recorded in the dictionary of the Slovenian literary language.

To quantitatively describe the global syntactic differences in the four age groups, we first computed the Zipf's exponent, the average length of the books and the unique word density. Results are presented in Fig. 2. It can be observed that the average values of the Zipf's exponents monotonically decrease with increasing AG and the differences are the most apparent in younger AG. This result suggests that the exponent converges and its value in the oldest AG (around −0.9) is comparable to those observed in other literary works [8, 22, 26, 31]. Moreover, the lengths of texts somehow expectedly increase with AG (see Fig. 2B). It should be noted that we checked whether there was any correlation between the length of the text and the Zipf's exponent and found no correlation between the two parameters. Apparently, texts

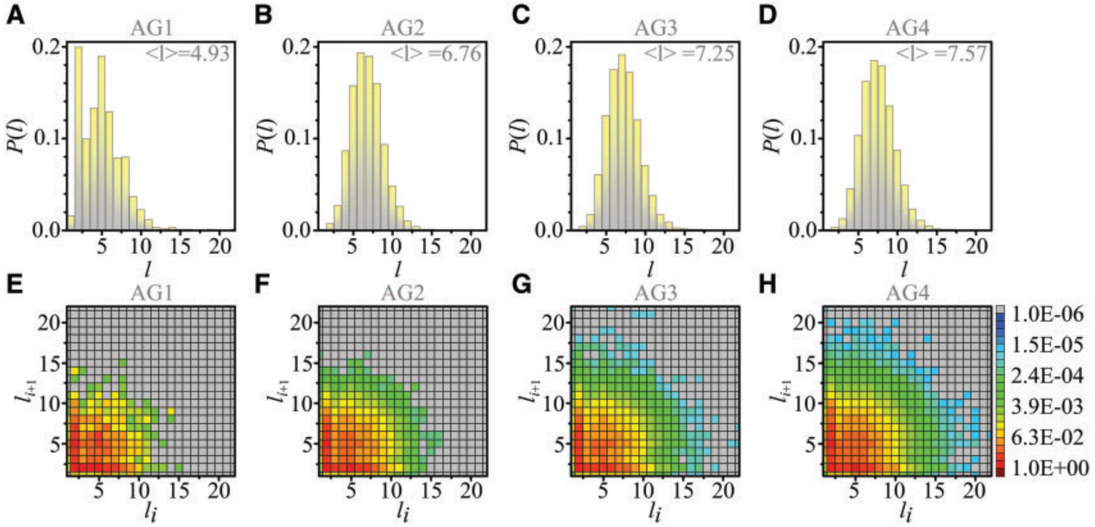

FIG. 3. Word length characteristics for different AGs. Distribution of word lengths (A–D) and 2D histogram showing how frequently two neighbouring words with a given length appear within a text (E–H). The average length l value for each subgroup is added to the graph (A–D).

for different AG differ in their global linguistic characteristics. A less negative exponent indicates less heterogeneity in the word frequency distribution.

It is also interesting, how the relationship between the number of unique words and the number of all words changes. Results are shown in Fig. 2C. In books, which are recommended for the first AG, the value of the unique word density parameter is around 0.5. This could be taken as an indicator, which the vocabulary acquisition is in the forefront. This in turn restricts a highly heterogeneous syntactic structure. In higher age groups, however, we can see that the density of unique words is significantly lower. In the last AG unique words represent only around 20% of all words. Since with every passing year the children's word treasure enriches, the need to learn new words to normally participate in social interactions through communication, decreases. Noteworthy, the most significant increase in the length of the texts is observed in later AG, as opposed to changes in the Zipf's exponent. It can thus be conducted that the style of writing (from a statistical point of view) for the oldest AG is rather matured and the focus is then placed to build up longer and more complex storylines, as addressed in more detail in the continuation.

Next, we focus on the lengths of words and the corresponding transitions between them. More precisely, we examine the distributions of lengths of words and the frequency of transitions between two neighbouring words of given lengths, separately for each AG. Results are shown in Fig. 3, and they highlight some differences among books for different AGs. In general, the average length of words increases for higher age groups. While the distributions for the highest three AGs have a very similar shape, the distribution in the lowest AG possesses one additional peak at the word length 2. This reflects the fact that the writing style for small children is different, more simple and with many conjunctions. Moreover, differences between AGs can also be detected in word sequencing. Figure 3E–H show the number of occurrences that of a word with a length l_i to be followed with a word of length l_{i+1}.

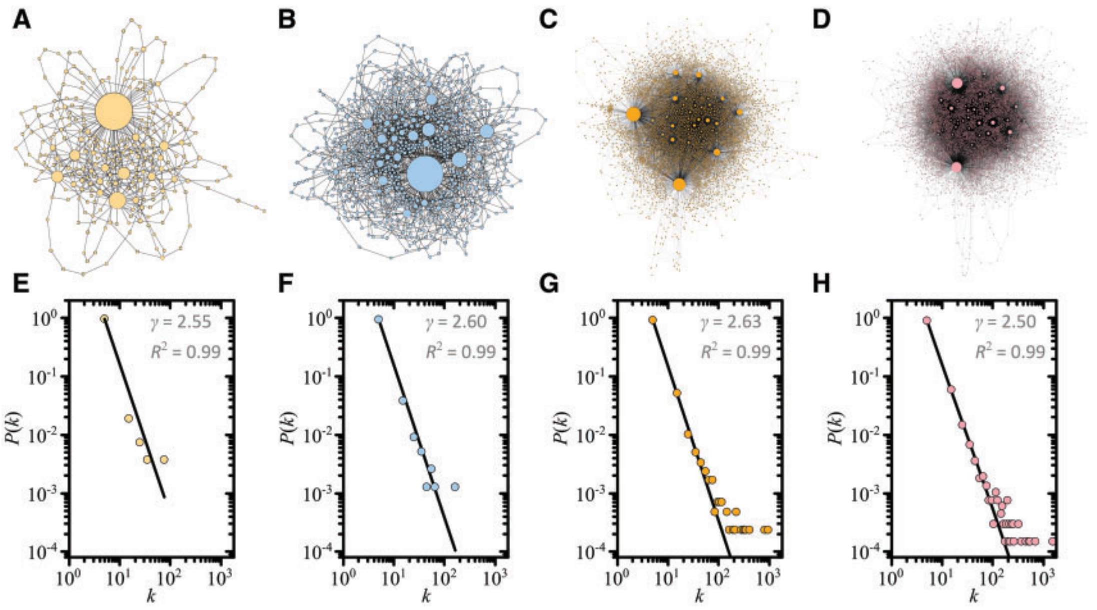

FIG. 4. Visualized language networks and degree distributions. Language networks for selected stories in individual age groups (A–D; the size of nodes is proportional to its degree) and the corresponding degree distributions (E–F).

For shorter words the transitions are rather similar between all AGs, but for the oldest two AGs more longer words as well as transitions between those longer words are detected, which reflects a more enriched vocabulary in the literature for oldest AG.

Next, we focus on the network driven approach and present a comprehensive analysis of the extracted word co-occurrence network characteristics. Figure 4 features four typical networks extracted from fables in different AGs and the corresponding degree distributions. Despite noticeable differences in networks sizes, the degree distribution exhibits a power-law behaviour in all four AGs with similar exponents. Apparently, some characteristics of syntactic patterns does not differ much between AGs, although the sizes of the investigated networks do not allow more precise assessments. Noteworthy, the observed scale-free property of language networks has previously been detected in several other settings and circumstances [23–31].

To get a more precise insight into the traits of the language and word interactions in different AGs, we calculated several network parameters and listed them in Table 2. Networks under study were obtained by combining all literary works in a given AG. The average degree, k, as well as the average clustering coefficient, C, increase with the recommended age of readers. This result goes in parallel with the decreasing tendency of unique word density. Namely, in higher AGs proportionally less unique words are used for longer texts and hence more word combinations are present, which in turn leads to more connections between individual words as well as to higher levels of cliquishness. For the same reasons, the networks also become less modular. Furthermore, on account on increasing number of connections, the average shortest path length progressively decreases with AG, despite the huge increase in network size. This in turn gives rise to the small-world topological features of the extracted word co-occurrence networks. Interestingly, despite changes in the average connectivity, the average shortest path length and network size, the diameter of the network is practically the same in all AGs.

TABLE 2 *The basic quantitative characteristics of the extracted language networks for all four AGs. The table contains the number of nodes N, average degree k; the average clustering coefficient C; the average shortest path length L; the modularity Q; small-worldness S and diameter D*

	N	k	C	L	Q	S	D
AG1	1017	4.26	0.16	3.40	0.50	44.5	9
AG2	9573	6.65	0.24	3.15	0.36	631.0	8
AG3	27 901	9.26	0.31	3.05	0.29	1629.9	9
AG4	68 676	11.4	0.37	2.95	0.27	3387.1	9

TABLE 3 *The basic quantitative characteristics of the extracted character networks for all four AGs. The table contains the number of nodes N, average degree k; the average clustering coefficient C; the average shortest path length L; the modularity Q and small-worldness S*

	N	k	C	L	Q	S
AG1	6	4.67	0.93	1.07	0.00	1.00
AG2	9	3.75	0.82	1.46	0.16	1.69
AG3	15	3.60	0.75	1.74	0.21	3.67
AG4	22	4.09	0.60	2.26	0.32	3.56

Lastly, we focus on the complexity of the storyline in fables for different AGs by assessing the interactions between characters. For each AG, we have selected one typical fable and constructed the corresponding sociograms. For all four extracted character networks, we have computed the community structure, the degree distribution and other network parameters. Results are presented in Fig. 5 and Table 3. As expected, the number of characters is higher in the texts for older AGs than in literary works for small children. It can also be nicely observed how social interactions between characters are gaining on complexity. While in the story from the first AG, there are a few interconnected characters, the novel for the oldest AG exhibits rather complex relationships, which bear similarities with real-life social networks: a heterogeneous degree distribution, an obvious community structure and small-worldness. Moreover, the main protagonists are the most connected nodes and they are typically poised in the central positions, linking different subgroups of characters [41].

4. Conclusion

In this article, a detailed statistical and topological analysis of 53 Slovene belles-lettres categorized into four different AGs has been performed. By merging all the texts into a single document the corresponding word frequency distribution follows the Zipf's law with the exponent value of −0.89, which is comparable to the exponent values calculated in other studies [16, 21, 49]. The average length of all words is 7.36, with the most commonly used words being the length of seven.

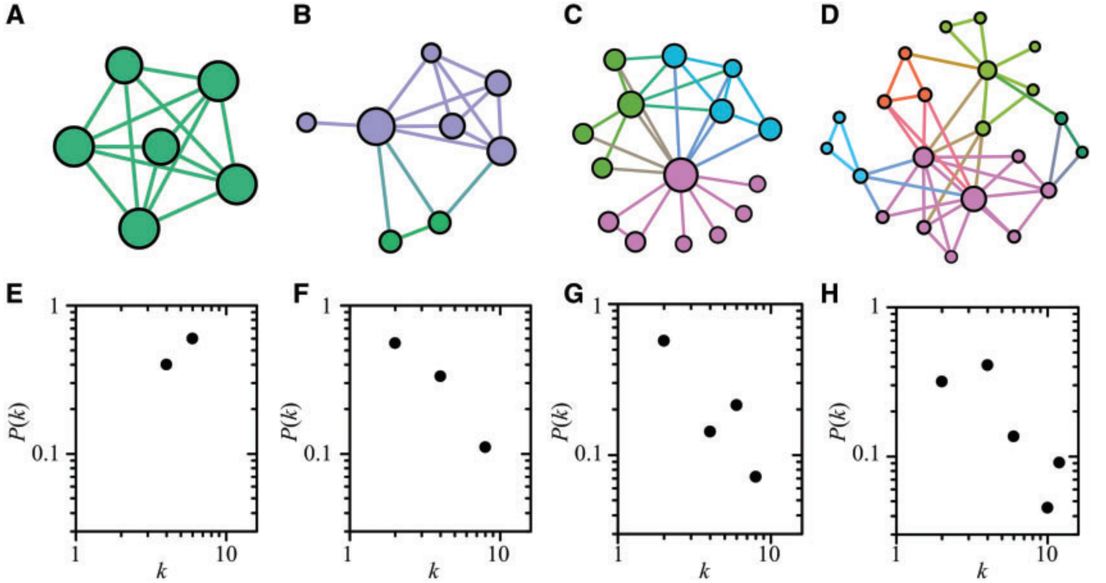

FIG. 5. Character networks from four stories for each AG. Sociograms for the fairies Muca copatarica AG1 (A), Pepelka AG2 (B), Kekec AG3 (C) and Pod svobodnim soncem AG4 (D) and the corresponding degree distributions (E–H). Colours in the sociograms (online version only) indicate the community structure and the size of nodes is proportional to its degree.

Within individual age groups texts have been merged and statistically evaluated by the Zipf exponent, the overall number of words, the unique word density, word length distribution and transitions of words with regard to their lengths. Texts intended for younger children (AG1) have the highest unique word density combined with a less negative Zipf exponent and are on average the shortest (see Fig. 2). With increasing age of the readers, the Zipf's exponent becomes increasingly more negative, increasingly more words are being used and reused, which in turns means a lower unique word density. The most pronounced change in the Zipf's exponent value is noticed between the first and second AG, where, seemingly, the style of writing changes significantly. On the other hand, the biggest change in the length of the texts occurs between the third and fourth AG. Moreover, with increasing AG the words are on average getting longer and more different transitions between words begin to emerge.

Classical statistical properties and Zipf's law provide only superficial understanding of the organization of language since relations between words are not taken into account. A set of those relations may be naturally represented as a network and by this means the connections between words can be explored to capture fundamental traits of language under a global picture. In our study, we constructed word co-occurrence networks of literary texts separately for different AGs. Our results indicate that the language networks in either cases exhibit a power-law behaviour with similar exponents (see Fig. 4). With increasing recommended age of readers, the level of small-worldness increases and the networks extracted from literary texts for older AGs are less modular, which reflects mostly the increasing numbers of possible transitions between words. Small-world topological features have been identified in several other syntactic networks [27–31] and such attributes are in general associated with efficient information transmission [50, 51]. Finally, we extracted the character networks from fables for different AGs. The complexity of interactions was found to increase with AG and especially in the latest AG the fictional social network was found to bear many similarities with real-life social networks such as small-worldness, a heterogeneous degree distribution and a community structure.

The herein presented findings are based solely on Slovene belles-lettres. Slovene, as a Slavic language, belongs to a group of inflectional languages, which are known for their rich morphology [52]. In addition, Slovene language is among the few languages that uses in addition to singular and plural also the dual form, which causes an even greater number of words. Another feature of Slavic languages is the category of verbal aspect [52]. Nevertheless, the role of language specifics and their effect on global and local topological features is an ongoing research topic tackled by morphological typology. In this vein, Liu and Xu [53] applied complex-network approaches to investigate the word form networks and the lemma networks of fifteen different languages. The computed topological features for individual languages highlight that only minor differences in the structural parameters are present, but are still sufficient to serve as a means of overall classification with similar accuracy as modern linguistic typological approaches [53, 54]. However, irrespective of the language under study, its structural characteristics such as near-Zipfian behaviour as well as the scale-free and small-world topological features of word networks, are universal [23, 27–31, 53, 54].

In sum, words are a good example on how simple elements that combine can form complex structures such as fables and novels. Inquiring the language as a network or a system with interconnected elements has turned out to be a powerful measure that is used in contemporary linguistics. In our study, we have shown that the utilization of such quantitative analyses and network based-approaches can successfully identify differences in literary texts for different recommended age of readers. This substantiates the ideas of using such approaches for automated classifications, not only in relation to recommended AGs, but possibly also in genre categorization or artistic merit evaluation.

Funding

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding, Nos. P3-0396, I0-0029-0552, and P1-0055, as well as research project No. J1-7009).

References

1. Cristelli, M., Batty, M. & Pietronero, L. (2012) There is more than a power law in Zipf. *Sci. Rep.*, **2**, 812.
2. Nowak, M. A., Plotkin, J. B. & Jansen, V. A. A. (2000) The evolution of syntactic communication. *Nature*, **404**, 495–498.
3. Nowak, M. A. & Krakauer, D. C. (1999) The evolution of language. *Proc. Natl. Acad. Sci.*, **96**, 8028–8033.
4. McCowan, B., Doyle, L. R., Jenkins, J. M. & Hanser, S. F. (2005) The appropriate use of Zipf's law in animal communication studies. *Anim. Behav.*, **69**, F1–F7.
5. Hurford, J. R. (2000) The evolution of the critical period for language acquisition. *Cognition*, **40**, 159–201.
6. Aitchison, J. (1996). *The Seeds of Speech: Language Origin and Evolution*. Cambridge: Cambridge University Press.
7. McCowan, B., Hanser, S. F. & Doyle, L. R. (1999) Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Anim. Behav.*, **57**, 409–419.
8. Zipf, G. K. (1950) Human behaviour and the principle of least effort. *Econ. J.*, **60**, 808–810.
9. Li, W. (1992) Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Trans. Inf. Theory*, **38**, 1842–1845.
10. Ferrer-i-Cancho, R. & Elvevåg, B. (2010) Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS One*, **5**, e9411.
11. Kilgarriff, A. (2005) Language is never, ever, ever, random. *Corpus Linguist. Linguist. Theory*, **1**, 263–275.

12. SAICHEV, A., MALEVERGNE, Y. & SORNETTE, D. (2010) *Theory of Zipf's Law and Beyond*. Berlin, Heidelberg: Springer Berlin Heidelberg.

13. Newman, M. (2005) Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.*, **46**, 323–351.

14. FERRER-I-CANCHO, R. & SOLÉ, R. V. (2003) Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci.*, **100**, 788–791.

15. FERRER-I-CANCHO, R. & SOLÉ, R. V. (2001) Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *J. Quant. Linguist.*, **8**, 165–173.

16. PIANTADOSI, S. T. (2014) Zipf's word frequency law in natural language: a critical review and future directions. *Psychon. Bull. Rev.*, **21**, 1112–1130.

17. KELLO, C. T., BROWN, G. D. A., FERRER-I-CANCHO, R., ET AL. (2010) Scaling laws in cognitive sciences. *Trends Cogn. Sci.*, **14**, 223–232.

18. BENTZ, C., KIELA, D., HILL, F., ET AL. (2014) Zipf's law and the grammar of languages: a quantitative study of Old and Modern English parallel texts. *Corpus Linguist. Linguist. Theory*, **10**, 175–211.

19. PETERSEN, A. M., TENENBAUM, J. N., HAVLIN, S., ET AL. (2012) Languages cool as they expand: allometric scaling and the decreasing need for new words. *Sci. Rep.*, **2**, 943.

20. MEHRI, A. & JAMAATI, M. (2017) Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. *Phys. Lett. A*, **381**, 2470–2477.

21. MCCOWAN, B., DOYLE, L. R. & HANSER, S. F. (2002) Using information theory to assess the diversity, complexity, and development of communicative repertoires. *J. Comp. Psychol.*, **116**, 166–172.

22. BAIXERIES, J., ELVEVÅG, B. & FERRER-I-CANCHO, R. (2013) The evolution of the exponent of Zipf's law in language ontogeny. *PLoS One*, **8**, e53227.

23. CONG, J. & LIU, H. (2014) Approaching human language with complex networks. *Phys. Life Rev.* **11**, 598–618.

24. SOLÉ, R. V., COROMINAS-MURTRA, B., VALVERDE, S. & STEELS, L. (2010) Language networks: their structure, function, and evolution. *Complexity*, **15**, 20–26.

25. BORGE-HOLTHOEFER, J. & ARENAS, A. (2010) Semantic networks: structure and dynamics. *Entropy*, **12**, 1264–302.

26. FERRER-I-CANCHO, R. & SOLÉ, R. V. (2001) The small world of human language. *Proc. R. Soc. B Biol. Sci.*, **268**, 2261–2265.

27. MASUCCI, A. P. & RODGERS, G. J. (2006) Network properties of written human language. *Phys. Rev. E*, **74**, 026102.

28. FERRER-I-CANCHO, R., SOLÉ, R. V. & KÖHLER, R. (2004) Patterns in syntactic dependency networks. *Phys. Rev. E*, **69**, 051915.

29. ZHOU, S., HU, G., ZHANG, Z., ET AL. (2008) An empirical study of Chinese language networks. *Phys. A*, **387**, 3039–3047.

30. LIU, H. (2008) The complexity of Chinese syntactic dependency networks. *Phys. A*, **387**, 3048–3058.

31. HOLOVATCH, Y. & PALCHYKOV, V. (2017). Complex networks of words in fables. *Maths Meets Myths: Quantitative Approaches to Ancient Narratives* (Kenna, R., MacCarron, M. & MacCarron P. eds), Understanding Complex Systems. Cham: Springer, pp. 159–175.

32. BECKAGE, N., SMITH, L., HILLS, T., ET AL. (2016) Small worlds and semantic network growth in typical and late talkers. *PLoS One*, **6**, e19348.

33. STEYVERS, M. & TENENBAUM, J. B. (2005) The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cogn. Sci.*, **29**, 41–78.

34. STELLA, M., BECKAGE, N. M., BREDE, M., ET AL. (2018) Multiplex model of mental lexicon reveals explosive learning in humans. *Sci. Rep.*, **8**, 2259.

35. DE JESUS HOLANDA, A., TORRES PISA, I., KINOUCHI, O., ET AL. (2004) Thesaurus as a complex network. *Phys. A*, **344**,530–536.

36. MOTTER, A. E., DE MOURA, A. P. S., LAI, Y.-C., ET AL. (2002) Topology of the conceptual network of language. *Phys. Rev. E*, **65**, 065102.

37. LIU, H. (2009) Statistical properties of Chinese semantic networks. *Sci. Bull.*, **54**, 2781–2785.

38. CHOI, Y.-M. & KIM, H.-J. (2007) A directed network of Greek and Roman mythology. *Phys. A*, **382**, 665–671.

39. Gleiser, P. M. (2007) How to become a superhero. *J. Stat. Mech. Theory. Exp.*, **2007**, P09020–P09020.

40. Mac Carron, P. & Kenna, R. (2012) Universal properties of mythological networks. *Europhys. Lett.*, **99**, 28002.

41. Beveridge, A. & Shan, J. (2016) Network of Thrones. *Math. Horizons*, **23**, 18–22.

42. Tan, M. S. A., Ujum, E. A. & Ratnavelu, K. (2017) Social network analysis of character interaction in the Stargate and Star Trek television series. *Int. J. Mod. Phys. C*, **28**, 1750017.

43. Das, D., Das, B., & Mahesh, K. (2016). A computational analysis of Mahabharata. *In: Proceedings of the 13th International Conference on Natural Language Processing* (Sharma, D. M., Sangal, R. & Singh, A. K. eds), Varanasi, India: NLP Association of India, pp. 219–228.

44. Liu, H. & Hu, F. (2008) What role does syntax play in a language network? *Europhys. Lett.*, **83**, 18002.

45. Barabási, A. L. (2016). *Network Science*. Cambridge, UK: Cambridge University Press.

46. Watts, D. J. & Strogatz, S. H. (1998) Collective dynamics of “small world” networks. *Nature*, **393**, 440–442.

47. Girvan, M. & Newman, M. E. J. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, **99**, 7821–7826.

48. Humphries, M. D. & Gurney, K. (2008) Network ‘Small-World-Ness’: a quantitative method for determining canonical network equivalence. *PLoS One*, **3**, e0002051.

49. Smith, R. (2012) Distinct word length frequencies: distributions and symbol entropies. *Glottometrics*, **23**, 7–22.

50. Barabási, A.-L. & Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

51. Estrada, E. (2012) *The Structure of Complex Networks: Theory and Applications*. New York: Oxford University Press.

52. Rotovnik, T., Maučec, M. S. & Kačič, Z. (2007) Large vocabulary continuous speech recognition of an inflected language using stems and endings. *Speech Commun.*, **49**, 437–452.

53. Liu, H. & Xu, C. (2011) Can syntactic networks indicate morphological complexity of a language? *Europhys. Lett.*, **93**, 28005.

54. Gao, Y., Liang, W., Shi, Y. & Huangd, Q. (2014) Comparison of directed and weighted co-occurrence networks of six languages. *Phys. A*, **393**, 579–589.