

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/301880096>

In narrative texts punctuation marks obey the same statistics as words

Article in *Information Sciences* · April 2016

DOI: 10.1016/j.ins.2016.09.051

CITATIONS

36

READS

836

4 authors:



[Andrzej Kulig](#)

Institute of Nuclear Physics, Polish Academy of Sciences, Polish Academy of Sciences

10 PUBLICATIONS 245 CITATIONS

[SEE PROFILE](#)



[Jarosław Kwapień](#)

Institute of Nuclear Physics, Polish Academy of Sciences, Polish Academy of Sciences

112 PUBLICATIONS 4,234 CITATIONS

[SEE PROFILE](#)



[Tomasz Stanisław](#)

Institute of Nuclear Physics, Polish Academy of Sciences, Polish Academy of Sciences

22 PUBLICATIONS 245 CITATIONS

[SEE PROFILE](#)



[Stanisław Drożdż](#)

Institute of Nuclear Physics and Cracow University of Technology

242 PUBLICATIONS 6,683 CITATIONS

[SEE PROFILE](#)

In narrative texts punctuation marks obey the same statistics as words

Andrzej Kulig^{1*}, Jarosław Kwapień¹, Tomasz Stanisławski¹, Stanisław Drożdż^{1,2}

¹ *Complex Systems Theory Department, Institute of Nuclear Physics, Polish Academy of Sciences, ul. Radzikowskiego 152, Kraków 31-342, Poland*

² *Faculty of Physics, Mathematics and Computer Science, Cracow University of Technology, ul. Warszawska 24, Kraków 31-155, Poland*

Abstract

From a grammar point of view, the role of punctuation marks in a sentence is formally defined and well understood. In semantic analysis punctuation plays also a crucial role as a method of avoiding ambiguity of the meaning. A different situation can be observed in the statistical analyses of language samples, where the decision on whether the punctuation marks should be considered or should be neglected is seen rather as arbitrary and at present it belongs to a researcher's preference. An objective of this work is to shed some light onto this problem by providing us with an answer to the question whether the punctuation marks may be treated as ordinary words and whether they should be included in any analysis of the word co-occurrences. We already know from our previous study [5] that full stops that determine the length of sentences are the main carrier of long-range correlations. Now we extend that study and analyze statistical properties of the most common punctuation marks in a few Indo-European languages, investigate their frequencies, and locate them accordingly in the Zipf rank-frequency plots as well as study their role in the word-adjacency networks. We show that, from a statistical viewpoint, the punctuation marks reveal properties that are qualitatively similar to the properties of the most frequent words like articles, conjunctions, pronouns, and prepositions. This refers to both the Zipfian analysis and the network analysis. Our results can be exploited in the computer-based analyses of large text corpora and be incorporated in the

*Corresponding author: andrzej.kulig@ifj.edu.pl

related automated systems. As a side result, we propose an efficient method of sampling the language corpora for a word-adjacency network analysis.

Keywords:

Word-adjacency networks, Complex networks, Word-frequency distribution

PACS: 89.75.-k, 89.75.Da, 89.75.Hc, 02.10.Ox

1. Introduction

Natural language is one of the most vivid examples of complex systems [15], where the term *more is different* [3] like no other succinctly defines its features. Indeed, the relatively small number of elementary items, the phonemes and letters, allow one to create more complex elements: the words. They form references to everything that a human can name and describe. However, the words alone do not constitute the whole essence of language and another complex entity is a prerequisite here: the sentence [5]. The sentential structure is a standard feature of almost all written languages. Only at this level the semantics in its whole richness and with a variety of carriers emerges: words, syntax, phrases, clauses, and punctuation in written language.

Statistical analyses of language samples that were carried out since over a century ago [6, 22] revealed the existence of laws that describe language quantitatively. Classical statistical study comprises, among others, the empirical word frequency distribution that is compared with the power-law model known as the Zipf law [23] and the functional relation between the length of a text and the number of unique words used to compose it, modelled by the Heaps law [8, 10, 11]. A relatively new approach is a description of language in the network formalism [4, 7, 9, 18, 19] that, among others, reveals that certain network representations of the lexical structure of texts (e.g. the word co-occurrence) belong to the scale-free class, similar to the semantic networks constructed based on the meaning of words [1, 2, 16].

Writing requires the use of punctuation; otherwise some expressions might be ambiguous and deceptive. Punctuation also allows one to denote separate logical units into which any compound message can be divided and be made easily comprehensible. From this perspective, the punctuation marks are something more than merely technical signs serving to allow a reader to comprehend the consecutive pieces of texts more easily. If put in between the words, they also acquire meaning and become meaningful not less than, for

example, some words playing mainly grammatical role as conjunctions and articles. For example, even though the full stops do not have clear phonetic expression, they define the length of sentences and thus they can influence a reader’s subjective perception of the message content: the speed of events, the richness, or multidimensionality of a described situation, etc. Our recent study shows additionally that punctuation carries long-range correlations in narrative texts [5]. This brings us more quantifiable evidence that punctuation, even though “silent”, is no less important than words.

Thus, it might seem intuitively natural to include such marks in any analysis, in which the ordinary words are considered: the rank-frequency, the word co-occurrence, and other types of the statistical analyses. It is sometimes done so in the engineering sciences like natural language processing due to practical reasons [12], but without any deeper linguistic justification. On the other hand, such an inclusion might not be recommended if the statistical properties of the punctuation marks were significantly different from the corresponding properties of the ordinary words as it would actually mean that the punctuation marks were something different than words. So, this issue appears to be rather a complex one. In order to resolve it, in this work we study the rank-frequency distributions and the word-adjacency networks in the corpora, in which the punctuation marks are treated as words, and compare the results for the punctuation marks with the results for the ordinary words. We argue that these results, which are complementary to the earlier ones published in [5], can provide one with indication on how to improve reliability of the statistical calculations based on large corpora of the written language samples that are performed by the automatic systems.

2. Data and methods

A literary form that is relatively the closest to the spoken language - prose - is expected to reflect the statistical properties of language. In order to analyze it, we selected a set of well-known novels written in one of six Indo-European languages belonging to the Germanic (English and German), Romance (French and Italian), and Slavic (Polish and Russian) language groups. Our selection criterion was the substantial length of each text sample, i.e., at least 5,000 sentences, which we have already verified to be sufficient for a statistical analysis [5]. The texts were downloaded from the Project Gutenberg website [21]. Apart from the individual texts, we also created 6 monolingual corpora by merging together at least 5 texts written in the

same language so that each corpus consisted of about one million words – a volume that was sufficient for our statistical analysis (see Appendix for a list of texts).

Some redundant words residing outside the sentence structure of texts (such as *chapter*, *part*, *epilogue*, etc.), footnotes, page numbers, and typographic marks (quotation marks, parentheses, etc.) were deleted. All standard abbreviations specific to a given language (like *Mrs.* and *Dr.* in English) were cleaned of dots and counted as separate words. The following marks were considered the full stops that end a sentence: dots, question marks, exclamation marks, and ellipses. For a comparative purpose, our analysis was extended to include commas (which were not considered the full stops, obviously). Our preliminary study was based on the frequency of word occurrence in a sample, which is a standard approach. It allowed us to check for possible statistical similarities between full stops or commas, and the ordinary words. It also aimed at testing whether these additional elements obey the well-known empirical Zipf law.

Next, in a word-adjacency network representation, where nodes represent words and connections represent the words' adjacent positions, the full stops and commas were taken into account like usual words. Doing so has practical importance for the consistency of the network creation process: otherwise there might be a problem whether the node representing a word ending a sentence and the node representing a word that starts the subsequent sentence may be connected to each other. On the one hand, such words are more loosely related semantically than the words within the same sentence are, but, on the other hand, leaving those nodes unconnected can lead to the formation of a disconnected network, for which many useful network measures cannot be well-defined. Identification of the punctuation marks as words thus allowed us to overcome this difficulty and to apply all the standard network measures effectively.

All calculations were performed in Mathematica and C++ environments independently. For better comparison between the corresponding results, all respective figures are shown in the same scale ranges.

3. Main results

3.1. Zipf analyses for language samples

The primary characteristics of natural language samples describing its quantitative structure is the Zipf distribution. It states that the probability

$P(R)$ of encountering the R th most frequent word scales according to $P(R) \sim R^{-\alpha}$ for $\alpha \approx 1$. There are different hypotheses on the origin of the Zipf law, with the principle of least effort [23] and the communication optimization [17] among them. It should be noted that this situation occurs only when a language sample is created in the unconstrained and spontaneous conditions. Existing aberrations from a power-law regime have appropriate justifications that have their source in an intellectual disability [20] or in sophisticated creative workshops [14].

After calculating the frequency of words, a set of words that are present in almost every sample is selected. As it turns out, for a sufficiently large sample they are always the words having grammatical functions. Regardless of the topics covered by a sample text, these words occupy the first ranks in the Zipf distribution. Additionally, we count the occurrence numbers of different punctuation marks in each sample and include them in the corresponding Zipf distributions as if they were ordinary words. The main plots in Fig. 1(a)-(f) show such distributions with distinguished punctuation marks (the special division): dot (#dot), question mark (#qu), exclamation mark (#ex), ellipsis (#ell), and comma (#com), respectively. In the insets to Fig. 1(a)-(f), all the marks that can end sentences are counted together as full stops (#fs).

Commas and the different types of full stops (except for ellipses) appear in the same region of the Zipf distribution where the highest-ranked words reside, i.e., the function words, like conjunctions (especially in the Slavic languages), articles (the Romance and Germanic languages), and prepositions. In all the considered languages, comma has $R = 1$, while the rank of dot is typically $R = 2$, except for Italian ($R = 3$) and English ($R = 5$). The question and exclamation marks have considerably lower ranks that vary among the languages but in general can be found in the interval $10 < R < 30$. Only ellipses can behave as lexical words with their ranks sometimes being lower than $R = 100$. For the general division, the unified full stop becomes the second most frequent object after comma in all languages except for English, where it occupies rank $R = 3$ (after comma and *the*). The most interesting observation regarding the Zipf plots is that all the punctuation marks in both divisions are placed firmly in the power-law regime together with the words. This means that, at least from this point of view, they are indistinguishable from the words.

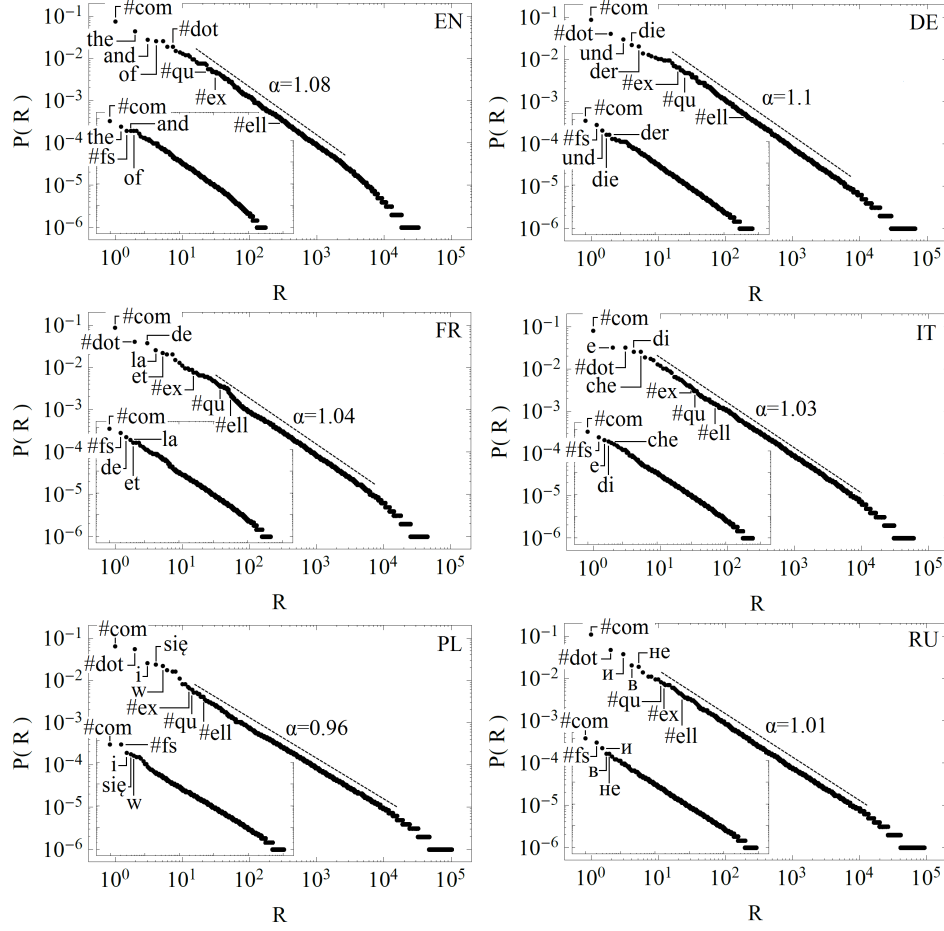


Figure 1: The word and punctuation-mark occurrence probability distributions for corpora representing different European languages: English (top left), German (top right), French (middle left), Italian (middle right), Polish (bottom left), and Russian (bottom right). Each language is represented by a corpus of length $s_c = 10^6$ words and punctuation marks created from a set of novels. For each language, the Zipfian scaling regime is indicated by a dashed line and a value of the related exponent α . (Main) Different punctuation marks are counted separately: comma (#com), dot (#dot), question mark (#qu), exclamation mark (#ex), and ellipsis (#ell). (Inset) All the punctuation marks that end sentences are counted together as full stops (#fs). In both panels the most frequent words are captioned.

3.2. Network properties for chosen words

Fig. 2 shows three stages of a word-adjacency network development. The network was created based on a growing sample of text of length s . The adopted representation allows us to check the adjacency relation between

words and punctuation marks. In Tab. 1 the chosen network parameters are shown for the corpora.

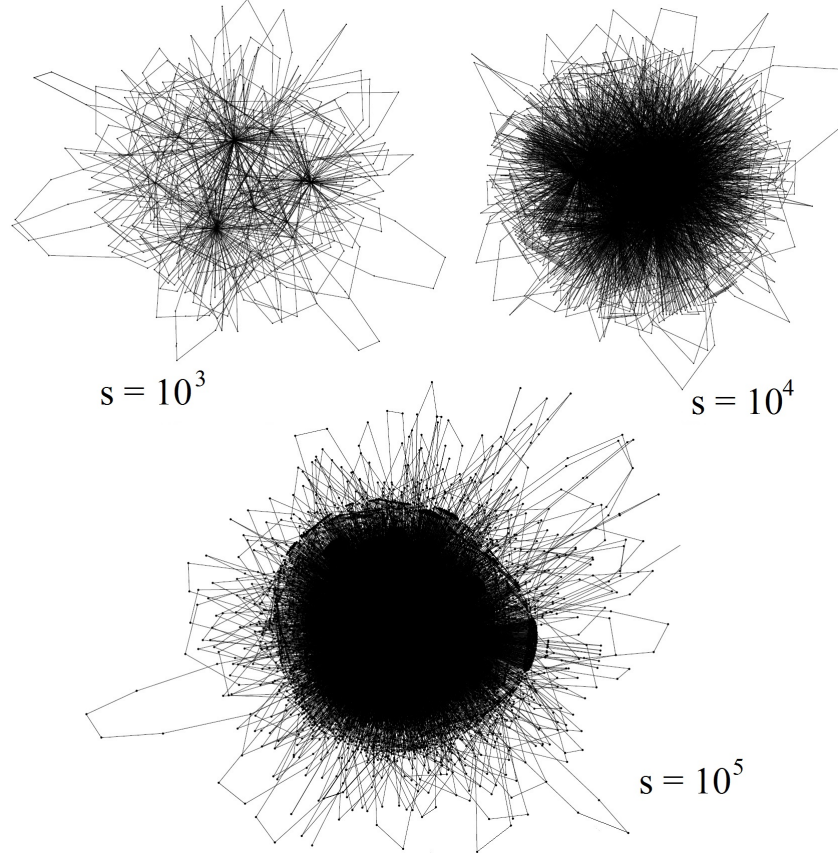


Figure 2: Typical forms of a growing word-adjacency network created from text samples of length $s = 10^3, 10^4, 10^5$ words.

	English	German	French	Italian	Polish	Russian
n	31784	65815	44785	60982	89990	91046
e	272501	373078	299897	374296	463219	421395

Table 1: Number of nodes n (vocabulary size) and unique edges e for word-adjacency network created based on monolingual corpora comprising $s = 10^6$ words. Since words were not lemmatized, the differences in n between the languages come predominantly from inflection.

A weighted work-adjacency network can be easily created from a text sample. The number of word co-occurrences may be understood as the weight of a connection between the respective nodes. The basic local parameter of the i th node is the number of edges attached to it, called a node degree k_i^w . It is roughly equal to doubled frequency f_i of the corresponding word in the sample. For a binary network, a node degree $k_i \equiv k_i^u$ refers to the number of unique connections from the i th node to other nodes, where f_i is the larger with respect to k_i , the more connections with other nodes this node has. In Fig. 3 the difference between f_i and k_i is shown for the 9 most frequent words, in a proper order starting from the left-hand side.

In English (Fig. 3(a)), these differences for all the considered words are substantial and roughly similar in size on logarithmic scale. This means that there exists a simple relation: $f_i \simeq a(i)k_i$ with $1/6 < a(i) < 1/3$. The most frequent English words often form 2-grams that are repeated many times throughout the corpus, which significantly lowers the degrees of the corresponding nodes. There is also no significant difference observed between comma, full stop and the other common words. In German, the articles: *die*, *der*, *den* are characterized by a small change between frequency and degree ($0.5 < a(i) < 1$). A different behavior is observed in the case of pronouns: *er*, *sie*, *ich*, where the changes between f_i and k_i are more significant ($0.3 < a(i) < 0.5$). The punctuation marks do not deviate from this picture, indicating that from this perspective they can be considered words.

In French, the pronouns: *le*, *la*, *il* show large differences, the prepositions: *de* and *a* show small differences, and the punctuation marks present moderate behavior (Fig. 3(b)). In Italian, all the considered objects except for full stop are characterized by small and steady difference between their frequency and degree. What is important, in contrast to the Germanic languages, there are comparable, rather small differences between f_i and k_i for the corresponding words in French and Italian. More significant differences between f_i and k_i are observed for Polish and Russian (Fig. 3(c)). The smallest difference is for a Polish conjunction *i* since this word does not have any special collocation with other words. In Russian all presented words are characterized by a significant $a(i)$, for pronouns the largest. The properties of the punctuation marks in both languages are alike.

For further calculations, two other local measures are used, that is, the average shortest-path length (ASPL) for a specific node ℓ_i and the local clustering coefficient C_i . ASPL for a node i refers to the average distance from a particular node to other nodes in the network and it is defined as

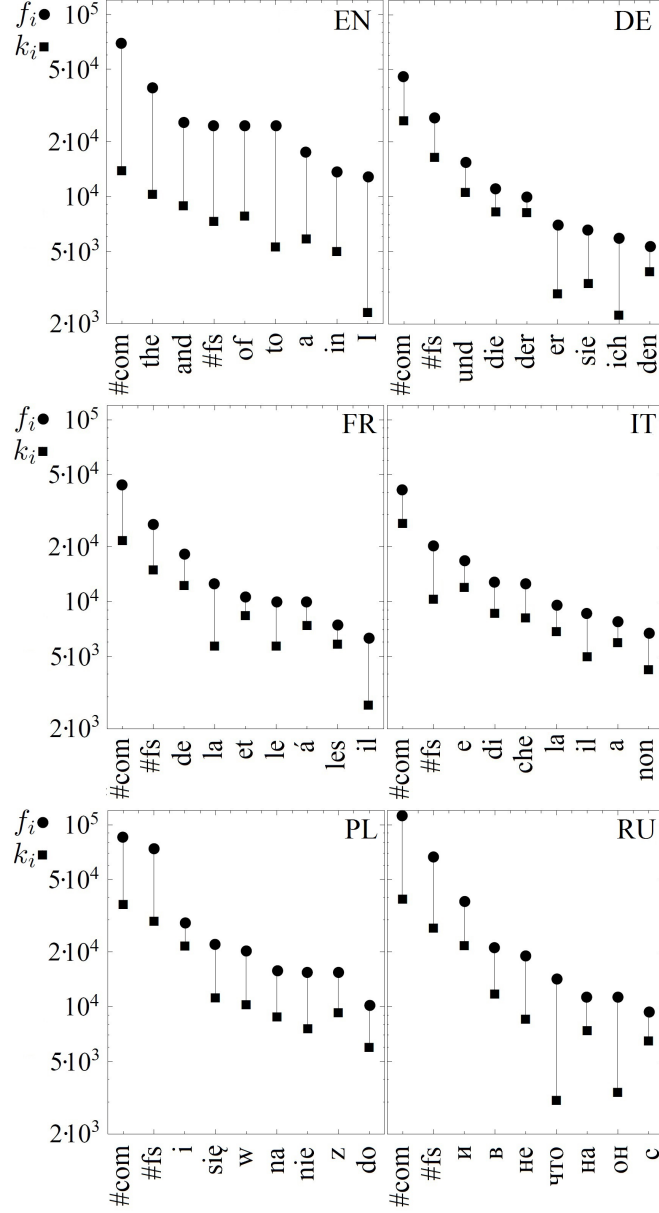


Figure 3: Difference between the frequency f_i of the most common words, full stops, and commas (circles) and the degree k_i of the respective nodes (squares) for the Germanic (top), Romance (middle), and Slavic (bottom) languages.

follows:

$$\ell_i = \frac{1}{n-1} \sum_j^n d(i, j), \quad (1)$$

where $d(i, j)$ denotes the shortest path (i.e., the one consisting of the minimal number of edges) between i and j , while n is the number of nodes in the network. The local clustering coefficient (LCC) for a node i is:

$$C_i = \frac{2e_i}{k_i(k_i - 1)}, \quad (2)$$

where e_i is the number of connections between direct neighbours of the i th node and k_i is its degree. This measure defines the density of links between direct neighbors of a given node and it can reveal membership of this node in a specific subset of strongly interconnected nodes [9].

In order to calculate ℓ_i and C_i , one has to note that both quantities depend on n [13]. This is because, according to the Heaps law, there is a non-linear dependency between the text size s and the vocabulary size n : $n \sim s^{\beta(s)}$ with $\beta(s)$ monotonically decreasing to zero for the infinitely long samples [8]. In result, with increasing sample size, the network becomes saturated gradually and tends to form almost a dense graph with only those edges missing that are forbidden by grammar. Therefore, typically ℓ_i decreases with increasing s , while C_i increases with s [13]. This effect can thus be observed also in the present study if we calculate both quantities for different values of s ($s \ll s_c$ in order to limit the calculation time).

Specifically, each monolingual corpus of length $s_c = 10^6$ was looped by connecting the last stop mark with the first word. Next, a substring of s words ($10^3 \leq s \leq 10^5$) was randomly chosen from the corpora and transformed into a word-adjacency network (by looping the corpora, it was always possible to create a substring of words of a given length if only $s \ll s_c$). This step was repeated $m = 100$ times giving a collection of m networks (we allowed for the substring overlapping since, for $s \ll s_c$, obtaining two identical substrings is unlikely). The network parameters ℓ_i and C_i were calculated for each network realization independently for the 9 most frequent words in each corpus and then their mean was also obtained: $\bar{\ell}_i = m^{-1} \sum_m \ell_i$ and $\bar{C}_i = m^{-1} \sum_m C_i$, respectively, together with its standard errors: σ_{ℓ_i} and σ_{C_i} .

The functional dependence of $\bar{\ell}_i(s)$ and $\bar{C}_i(s)$ for the most common English words is presented in Fig.4(a) and Fig.4(b), respectively. For the other languages considered here both plots look qualitatively similar except for that different words can be listed in each case. It is interesting to note that $\ell_i(s)$ for #fs and $C_i(s)$ for #com do not differ much from their counterparts

representing the ordinary words, $\ell_i(s)$ for comma is distinguished by exceptionally small values while preserving the monotonically decreasing shape of ASPL for the other objects, and only $C_i(s)$ for full stops reveals rather a different behaviour with the weakest variability. The results obtained for all the 6 languages are summarized in Fig. 5(a)-(f) in a form of scatter plots ℓ_i vs. C_i for the medium sample size of $s = 10^4$.

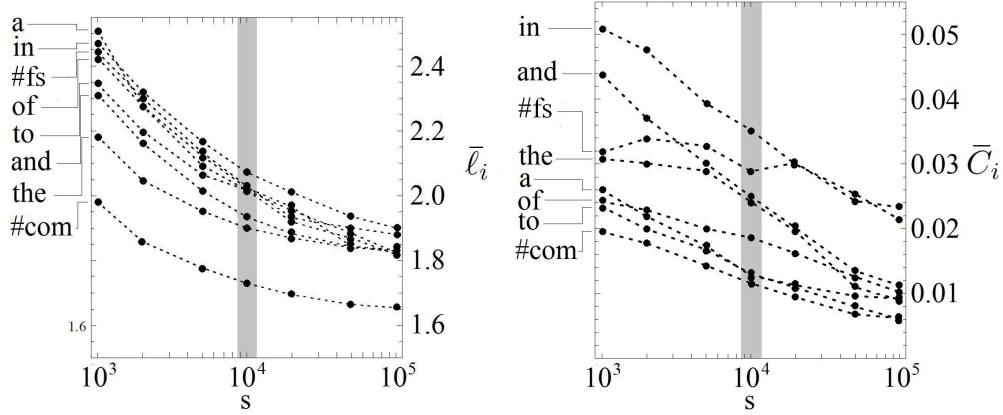


Figure 4: The word-specific average shortest path length $\bar{\ell}_i$ (left) and the local clustering coefficient \bar{C}_i (right) averaged over different text samples as functions of the sample size s for the most frequent English words, full stops ($\#fs$), and commas ($\#com$). The shaded region indicates the value of $s = 10^4$ used for creating Fig. 5.

The standard errors determined for ℓ_i and C_i are typically so small that they do not differ much from the symbol size in Fig. 5. On the one hand, owing to the ASPL definition, in each case the value of ℓ_i is negatively correlated with a node degree k_i ; on the other hand, LCC does not show any correlation with k_i . Typically, full stop and comma have rather low values of ASPL. Only the Italian and English full stops have larger value of ℓ_i that may originate from the significant differences between $f_{\#fs}$ and $k_{\#fs}$, see Fig. 3(a) and Fig. 3(b). Except for English, full stops have C_i that assumes its lowest values. Among the considered words, the most distinguished one is the English pronoun *I* with a significant variability of both ℓ_i and C_i among the individual sample networks. In terms of similarities among the Germanic languages, the pronouns *I* and *ich*, which have the same meaning, have also comparable mean values of ASPL and LCC. Moreover, in German some words having similar meaning and function (e.g. *die* and *der* or *sie* and *er*) have also similar coordinates in the corresponding scatter plot.

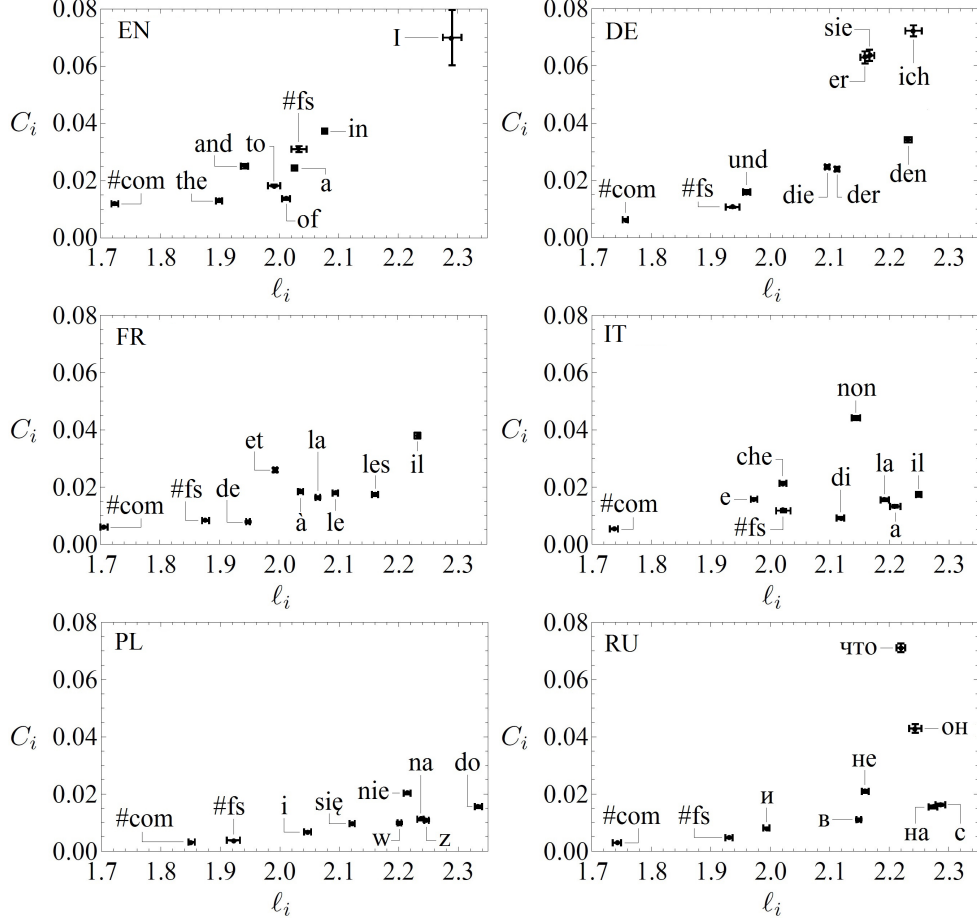


Figure 5: Scatter plots of the word-specific average shortest-path length $\bar{\ell}_i$ and the local clustering coefficient \bar{C}_i for the most frequent words, full stops #fs, and commas #com in six different European languages: English (top left), German (top right), French (middle left), Italian (middle right), Polish (bottom left), and Russian (bottom right). Error bars denote standard deviations calculated from 100 independent text samples.

The above results prove that the basic network properties for the most common words and the punctuation marks are similar. Now we consider another property of nodes, i.e., the indicators how important for the network structure their presence is. In other words, we study how the removing of particular nodes can impact the overall network structure expressed in terms of the global network measures. We look at three such measures: the average shortest path length: $L = \sum_i \ell_i$, the global clustering coefficient: $C = \sum_i C_i$,

and the global assortativity coefficient r :

$$r = \frac{\sum_{ij}(\delta_{ij} - \frac{k_i k_j}{2e})}{\sum_{ij}(k_i \delta_{ij} - \frac{k_i k_j}{2e})}, \quad (3)$$

where e is the number of edges in the network and δ_{ij} equals 1 if there is an edge between the nodes i and j or 0 otherwise. We now consider the nodes representing full stop and the 9 highest-ranked words in each text sample. Unlike before, we now do not consider comma, however, because the highest frequency of its occurrences in almost all text samples obviously results in the highest impact of the related node's removal.

For different text samples (novels), we compare values of $L(R)/\ln n$ (Fig. 6), $C(R)$ (Fig. 7), and $r(R)$ (Fig. 8) calculated for a complete network with all the nodes present (denoted by the abscissa $R = 0$) and for an incomplete network obtained by removing a given highly ranked node according to the Zipf distribution ($1 \leq R \leq 10$). The reason for dividing $L(R)$ by $\ln n$ (which is equal to L for the random Erdős-Rényi networks) is that it is the only measure among the ones considered here that depends on n , so without this normalization ASPLs for different novels could not be compared.

In each case by removing one of the highly connected nodes, ASPL becomes longer than for the complete network and this is not surprising since the network loses one of its hubs. This increase of $L(R)$ is different for different ranks and different novels but typically it does not exceed 1% for the highest ranks. Roughly, a rule is that for the larger R the change in $L(R)/\ln n$ is smaller, but this is true only statistically (for a particular novel there might be some exceptions). This rule comes from the fact that in the word-adjacency networks removing a strong hub is more destructive for the network than removing some less connected node. $L/\ln n$ varies among the novels also if we compare its values for the same objects (as a word or punctuation mark can occupy different ranks in different text samples).

The global clustering coefficient $C(R)$ and assortativity $r(R)$ present a more variable behaviour after removing a node as they can either increase, remain stable, or decrease. This behaviour obviously depends on a contribution of each particular node to C and r for $R = 0$, but a statistical rule is that without the most connected nodes like *#fs* and *and* (or its foreign counterparts like *et*, *und*, etc.), both C and r increase, which mean that the networks become more clustered and less disassortative. This can be explained by an observation that full stops (especially) and conjunctions can

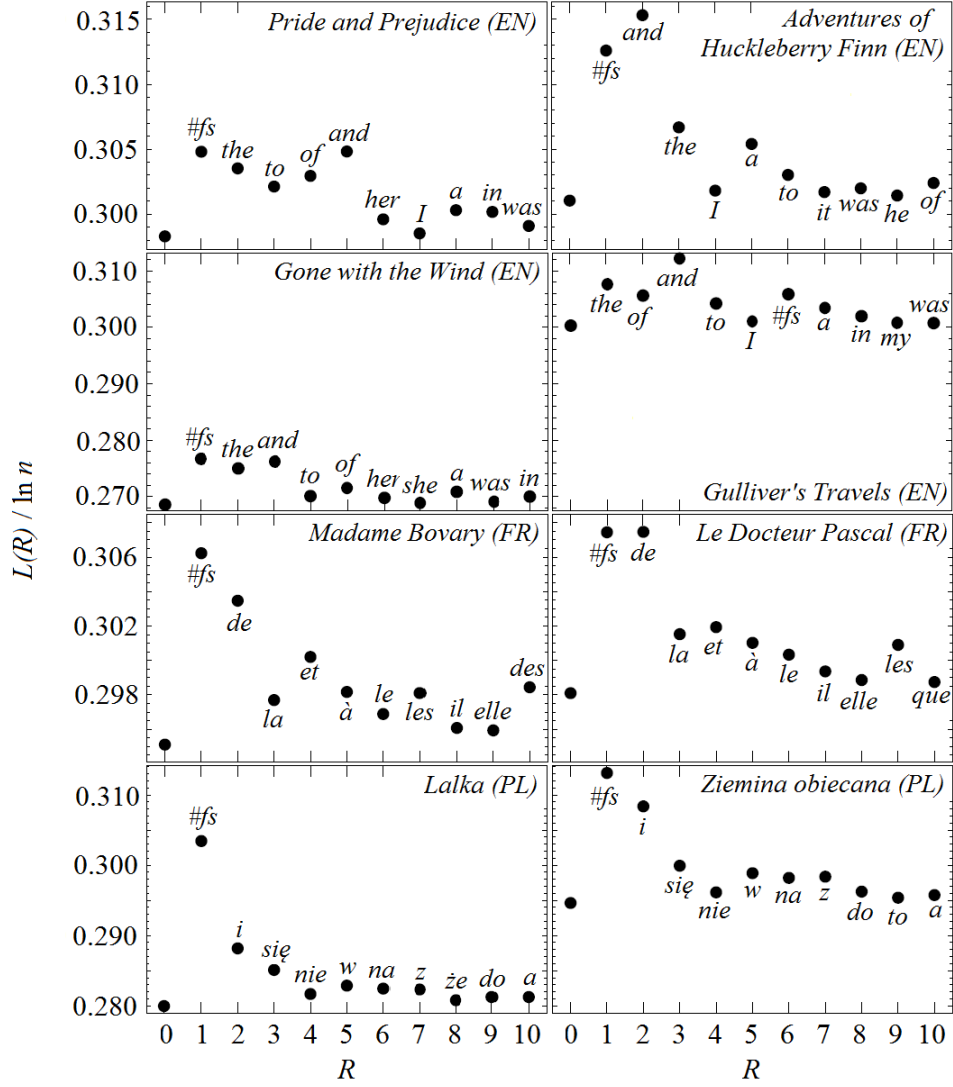


Figure 6: The average shortest path length $L(R)$ normalized to $\ln n$ for networks representing different novels. For each novel, the rank $R = 0$ denotes the complete network with all the n nodes, while the lower ranks $1 \leq R \leq 10$ denote the incomplete networks with $n - 1$ nodes obtained by removing a node corresponding to a word ranked R in the Zipf distribution for this novel.

mediate words that are placed within different sentences, different sentence parts or different phrases (therefore they can be neighbours of almost any word). Thus, they may link groups of words that are clustered together inside

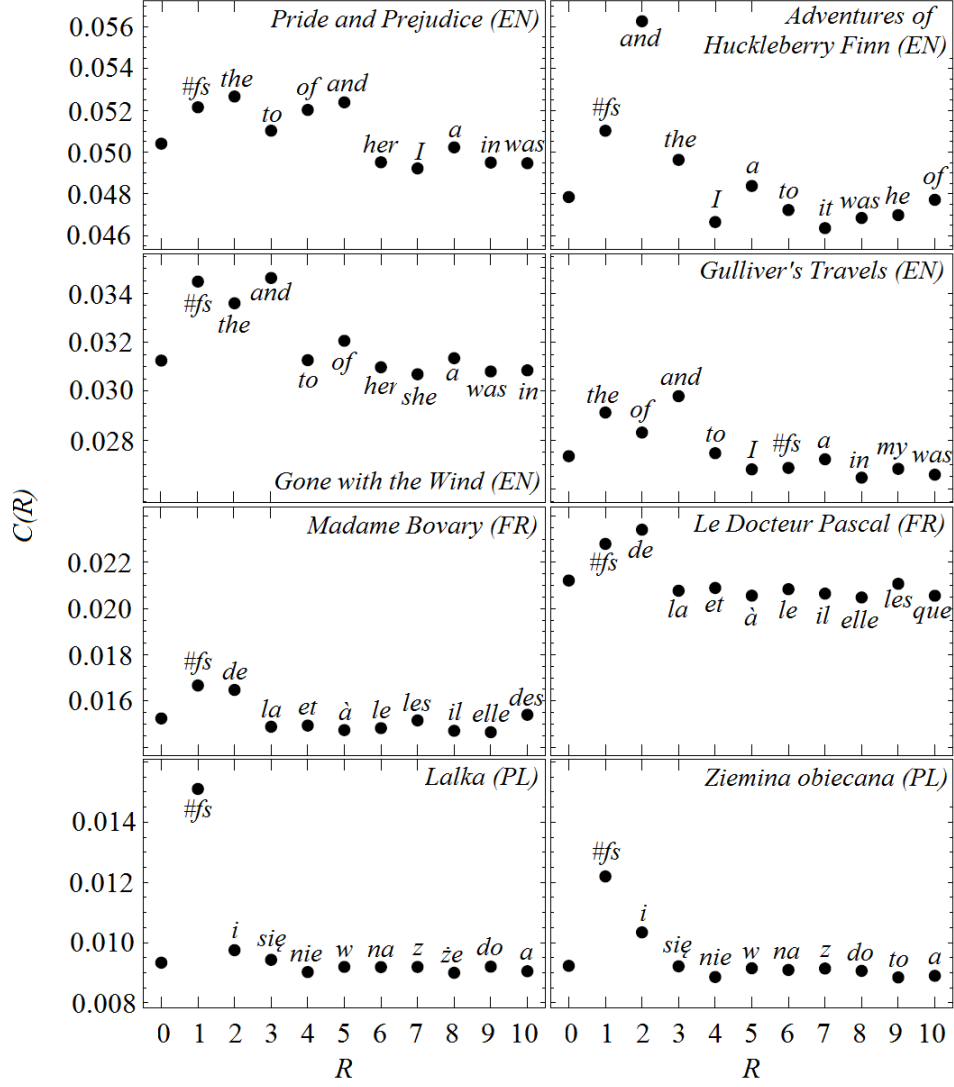


Figure 7: The global clustering coefficient $C(R)$ for networks representing different novels. For each novel, the rank $R = 0$ denotes the complete network with all the n nodes, while the lower ranks $1 \leq R \leq 10$ denote the incomplete networks with $n - 1$ nodes obtained by removing a node corresponding to a word ranked R in the Zipf distribution for this novel.

each group via their strong semantical relations, but that are less clustered between the groups due to their larger semantical distance. Their removal can thus increase GCC of the network. Other words that do not link sen-

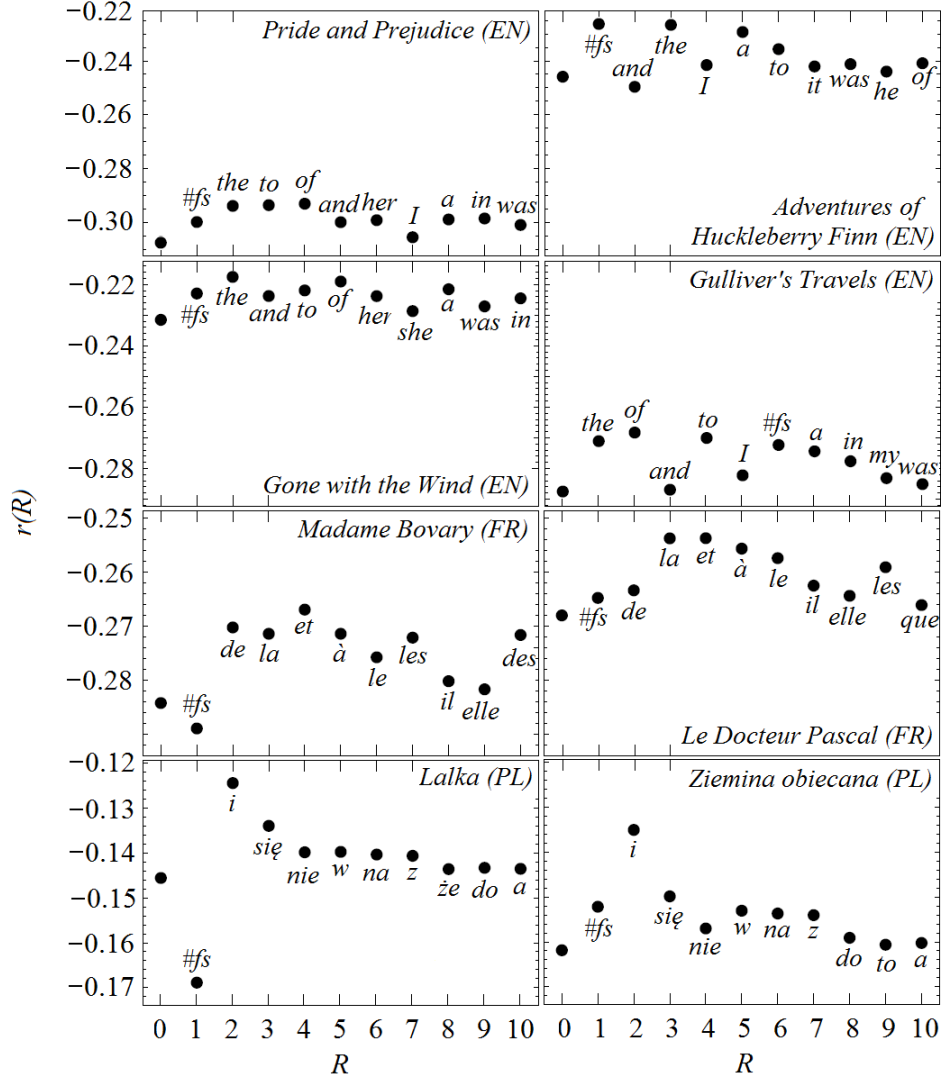


Figure 8: The global assortativity index $r(R)$ for networks representing different novels. For each novel, the rank $R = 0$ denotes the complete network with all the n nodes, while the lower ranks $1 \leq R \leq 10$ denote the incomplete networks with $n - 1$ nodes obtained by removing a node corresponding to a word ranked R in the Zipf distribution for this novel.

tences or phrases reside inside these structures and can be therefore part of clusters, so their removal decreases the overall network clustering. As regards the assortativity index $r(R)$, the hubs in the word-adjacency networks are typically connected in a disassortative manner, so after their disconnection

from the rest of the network, the overall assortativity index can increase. Of course, since this is only a statistical observation, particular cases may show different behaviour.

4. Conclusion

Punctuation marks are among the most common objects in written language. They do not play purely grammatical roles, but they also carry some semantic load, similar to such words like articles, conjunctions, and prepositions. This opens space for putting a question whether punctuation marks may be included in any lexical analysis on par with the ordinary words. In this work we addressed this question by comparing the statistical properties of punctuation marks and words using two approaches. We observed that punctuation marks locate themselves exactly on or in a close vicinity of the power-law Zipfian regime as if they were ordinary words. We drew the same conclusion from an analysis of the word-adjacency networks, in which words, full stops (the aggregated sentence-ending punctuation marks), and commas were considered nodes. In such networks, punctuation marks play a role of hubs together with the most frequent words. Despite some minor, quantitative-only differences, topology of such networks and their growth is similar from the perspective of punctuation marks and from the perspective of words. Quantitatively, it is expressed by the node-specific average shortest path length, the local clustering coefficient, the local assortativity, and their global counterparts. These results are qualitatively invariant under language change even for the languages belonging to different Indo-European groups. Regarding the quantitative viewpoint, we do observe certain systematic differences of the network properties between different text samples (including different languages), but considering them here is beyond the scope of this work. A related study will be presented and discussed elsewhere.

By taking all these outcomes into consideration, the principal conclusion from this study is that punctuation marks are almost indistinguishable from other most common words if one investigates their statistical properties. Since the punctuation marks have non-neglectable meaning also, we advocate their inclusion in any type of the word-occurrence and the word-adjacency analysis making it to be more complete.

5. Appendix

The books used in our analysis (asterisks denote the corpora-forming books):

English: George Orwell *1984**, Mark Twain *Life on the Mississippi**, *Adventures of Huckleberry Finn*, Jane Austen *Pride and Prejudice**, James Joyce *Ulysses**, Jonathan Swift *Gulliver's Travels**, Margaret Mitchell *Gone with the Wind*.

German: Friedrich Nietzsche *Also sprach Zarathustra**, Franz Kafka *Der Process**, Heinrich Mann *Der Untertan**, Thomas Mann *Der Zauberberg**, Christiane Vera Felscherinow *Wir Kinder vom Bahnhof Zoo**.

French: Alexandre Dumas *Ange Pitou**, Albert Camus *La Peste**, Émile Zola *La Terre**, *Le Docteur Pascal*, Gustave Flaubert *Madame Bovary**, Gaston Leroux *Le Fantôme de L'Opéra**.

Italian: Umberto Eco *Il pendolo di Foucault**, Gabriele d'Annunzio *Trionfo della morte**, Giambattista Bazzoni *Falco della Rupe o la guerra di Musso**, Luigi Capuana *Giacinta**, Tullio Avoledo *Le Radici del Cielo**.

Polish: Gustaw Herling-Grudziński *Inny świat**, Karol Olgierd Borchardt *Znaczy Kapitan**, Walery Łoziński *Zaklęty dwór**, Stefan Żeromski *Przedwiośnie**, Władysław Reymont *Ziemia obiecana**, Bolesław Prus *Lalka*.

Russian: Lev Tolstoy *Анна Каренина (Anna Karenina)**, *Война и мир (War and Peace)**, *Воскресение (Resurrection)**, Fyodor Dostoyevsky *Бесы (Demons)**, *Братья Карамазовы (The Brothers Karamazov)**.

References

- [1] D.R. Amancio, O.N. Oliveira Jr, L.D.F. Costa, Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts, *Phys. A* 391 (2012) 4406-4419.
- [2] D.R. Amancio, A complex network approach to stylometry, *PLoS ONE* 10 (2015) e0136076.
- [3] P.W. Anderson, More is different, *Science* 177 (1972) 393-396.
- [4] S.N. Dorogovtsev, J.F.F. Mendes, Language as an evolving word web, *Proc. R. Soc. Lond. B: Biol. Sci.* 268 (2001) 2603-2606.

- [5] S. Drożdż, P. Oświęcimka, A. Kulig, J. Kwapien, K. Bazarnik, I. Grabska-Gradzińska, J. Rybicki, M. Stanuszek, Quantifying origin and character of long-range correlations in narrative texts, *Inf. Sci.* 331 (2016) 32–44.
- [6] J.-B. Estoup, *Gammes sténographiques. Methodes et exercices pour l’acquisition de la vitesse*, Institut Sténographique de France, 1916.
- [7] R. Ferrer-i-Cancho, R.V. Solé, The small world of human language, *Proc. R. Soc. Lond. B: Biol. Sci.* 268 (2001) 2261–2265.
- [8] M. Gerlach, E.G. Altmann, Stochastic model for the vocabulary growth in natural languages, *Phys. Rev. X* 3 (2013) 021006.
- [9] I. Grabska-Gradzińska, A. Kulig, J. Kwapien, S. Drożdż, Complex network analysis of literary and scientific texts, *Int. J. Mod. Phys. C* 23 (2012) 1250051.
- [10] H.S. Heaps, *Information Retrieval: Computational and Theoretical Aspects*, Academic Press, Orlando, 1978.
- [11] G. Herdan, *Type-token Mathematics. A Textbook of Mathematical Linguistics*, Mouton, ’s-Gravenhage, 1960.
- [12] A. Kao, S.R. Poteet, *Natural language processing and text mining*, Springer Science & Business Media, Berlin, 2007.
- [13] A. Kulig, S. Drożdż, J. Kwapien, P. Oświęcimka, Modeling the average shortest-path length in growth of word-adjacency networks, *Phys. Rev. E* 91 (2015) 032810.
- [14] J. Kwapien, S. Drożdż, A. Orczyk, Linguistic complexity: English vs. Polish, text vs. corpus, *Acta Phys. Pol. A* 117 (2010) 716–720.
- [15] J. Kwapien, S. Drożdż, Physical approach to complex systems, *Phys. Rep.* 515 (2012) 115–226.
- [16] H. Liu, Statistical properties of Chinese semantic networks, *Chin. Sci. Bull.* 54 (2009) 2781–2785.

- [17] B.B. Mandelbrot, An information theory of the statistical structure of language, in: W. Jackson (ed.), Communication Theory, pp. 503-512, Academic Press, New York, 1953.
- [18] M. Markosova, Network model of human language, Phys. A 387 (2008) 661-666.
- [19] A.P. Masucci, G.J. Rodgers, Network properties of written human language, Phys. Rev. E 74 (2006) 026102.
- [20] W. Piotrowska, X. Piotrowska, Statistical parameters in pathological text, J. Quant. Ling. 11 (2004) 133-140.
- [21] The Project Gutenberg website, *www.gutenberg.org*.
- [22] G.K. Zipf, Selective Studies and the Principle of Relative Frequency in Language, MIT Press, Cambridge, 1932.
- [23] G.K. Zipf, Human behavior and the principle of least effort, Addison-Wesley, Cambridge, 1949.