Information Sciences 375 (2017) 98–113

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

In narrative texts punctuation marks obey the same statistics as words

Andrzej Kulig[a], Jarosław Kwapień[a], Tomasz Stanisz[a], Stanisław Drożdż[a,b,*]

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

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

ARTICLE INFO

Article history:
Received 1 April 2016
Revised 25 August 2016
Accepted 22 September 2016
Available online 22 September 2016

Keywords:
Punctuation
Word-adjacency networks
Complex networks
Word-frequency distribution

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 (S. Drożdż et al., Inf. Sci. 331 (2016) 32-44) that full stops that determine the length of sentences are the main carrier of long-range correlations. Now we extend that study and analyse 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. By adding the punctuation marks to the Zipf plots, we also show that these plots that are normally described by the Zipf–Mandelbrot distribution largely restore the power-law Zipfian behaviour for the most frequent items.

Our results indicate that the punctuation marks can fruitfully be considered in the linguistic studies as their inclusion effectively extends dimensionality of an analysis and, therefore, it opens more space for possible manifestation of some previously unobserved effects.

1. Introduction

Natural language is one of the most vivid examples of complex systems [18], where the term *more is different* [4] 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

* Corresponding author.
E-mail address: stanislaw.drozdz@ifj.edu.pl (S. Drożdż).

http://dx.doi.org/10.1016/j.ins.2016.09.051
0020-0255/

and describe. However, the words alone do not constitute the whole essence of language and another complex entity is a prerequisite here: the sentence [8]. 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 [9,27] 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 [28] or its generalized form known as the Zipf–Mandelbrot law [21,24], 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 [11,13,14]. A relatively new approach is a description of language in the network formalism [6,10,12,22,23] 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 [2,3,19].

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. 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 descriptive complexity of a given situation, etc. Our recent study shows additionally that punctuation carries long-range correlations in narrative texts [8]. 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 [5]. It is sometimes done so in the engineering sciences like natural language processing due to practical reasons [15], 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 [8], can provide one with indication on how to improve reliability of the statistical calculations based on large corpora of the written language samples.

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 analyse 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 5000 sentences, which we have already verified to be sufficient for a statistical analysis [8]. The texts were downloaded from the Project Gutenberg website [26]. 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. Apart from the full stops, our analysis also included commas, colons, and semicolons.

Moreover, the notion of the punctuation marks may be generalized in such a way that it includes new chapters, new parts, and new paragraphs (that are recognized as the separators stronger than a full stop), as well as new lines (that may further be divided into: comma-new line, colon-new line, etc.). While the division into parts is too sparse to be meaningful in our analysis and the localization of all new paragraphs and new lines is too demanding to be easily done here, we extended our analysis over the chapters. In each text we found the places, in which new chapters begin, and introduced them into the texts as an additional punctuation mark (denoted as #chap). We preferred not to consider any specific word as a separator in this context, because different ways of denoting new chapters are used in different texts: the word "chapter", the Roman or the Hindu-Arabic numerals, the asterisks, or even just the voids. One issue should be kept in mind, however. While the standard punctuation can be viewed as an inherent part of the natural language that helps one to understand the message, the division of texts into paragraphs, chapters, and parts is purely a writing technique not necessary from the point of view of the language organization.

Our first analysis 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 the punctuation marks 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 punctuation marks 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 Rth most frequent word scales according to $P(R) \sim R^{-\alpha}$ for $\alpha \approx 1$. The Zipfian scaling in its original formulation holds for the majority of ranks except for a few highest ones, where the power law breaks and the corresponding plots are deflected towards lower frequencies than those expected from the pure power law. Therefore a better agreement with the empirical data one can obtain using the so-called Zipf–Mandelbrot law (shifted power-law):

$$P(R) \sim (R + c)^{-\alpha}, \tag{1}$$

where c is the parameter responsible for the above-mentioned deflection. There are different hypotheses on the origin of the Zipf law, with the principle of least effort [28] and the communication optimization [20] among them. It should be noted that this situation occurs only if 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 [25] or in sophisticated creative workshops [17].

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 show such distributions with distinguished punctuation marks (the special division): dot (#dot), question mark (#qu), exclamation mark (#ex), ellipsis (#ell), semicolon (#scol), colon (#col), comma (#com), and new chapter (#chap). In the insets to Fig. 1, 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 and English ($R = 3$). The question and exclamation marks as well as semicolons and colons have considerably lower ranks that vary among the languages but in general can be found in the interval $10 < R < 30$ (#qu and #ex) and in the interval $10 < R < 50$ (#scol and #col). Ellipses can behave as lexical words with their ranks sometimes being lower than $R = 100$. This refers even more to the new chapter marks whose frequency varies strongly from text to text and their rank can be as low as $R \approx 1000$ for particular books. 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 plots is that all the punctuation marks in both divisions are placed together with the regular words in the regime that is close to a power-law. This means that adding the punctuation marks to the Zipf analysis results in a substantial improvement of the scaling of the rank-frequency plots in that part ($R < 10$) that in a standard analysis deviates from a power-law towards the lower frequencies and that is described by the Zipf–Mandelbrot distribution. From this point of view, the punctuation marks act towards restoring of the Zipf distribution. This effect can be seen in Fig. 1, where the Zipfian power law fitted within the range $[10^1, 10^4]$ is geometrically extended over the highest ranks. For all the languages the corresponding points are closer to the power law and for French, Polish, and Russian they are placed exactly in the scaling regime. For a comparison, in Fig. 2 we present analogous Zipf plots for two texts where the punctuation differs from the standard pattern (a lack of the sentence structure of the texts). However, except for the distant location or even the absence of #dot, the overall statistical properties of the remaining punctuation marks are normal.

To express the above observation in a quantitative form, we fit the Zipf–Mandelbrot (Eq. (1)) distribution to the rank-frequency plots constructed for words and for words together with the punctuation marks and estimate the corresponding values of the parameter c responsible for a deflection from the pure power law ($c = 0$). Fig. 3 shows such fits for all the considered languages. In each case, the inclusion of the punctuation marks results in the significantly lower values of c than those in the case, in which only the words are considered, with the strength of this decrease depending on a language. It

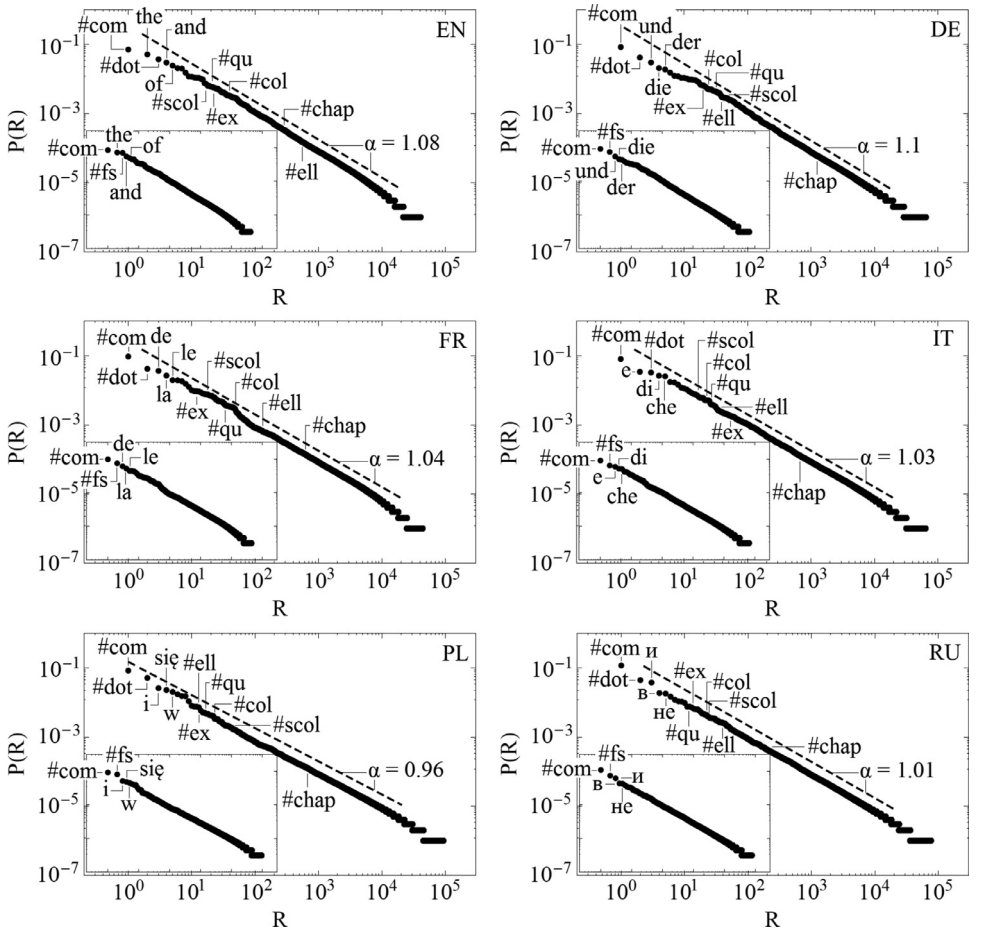

Fig. 1. The word and punctuation-mark occurrence probability distributions for corpora representing different European languages: English (top left), German (top right), French (middle left), Italian (middle right), Polish (bottom left), and Russian (bottom right). Each language is represented by a corpus of length $s_c \approx 10^6$ words and punctuation marks created from a set of novels. For each language, dashed lines represent the Zipfian power law fitted within the range $[10^1, 10^4]$ (and extended over the range $R < 10$) 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), ellipsis (#ell), semicolon (#scol), colon (#col), and new chapter (#chap). (Inset) All the punctuation marks that end sentences are counted together as full stops (#fs). In both panels the most frequent words are captioned.

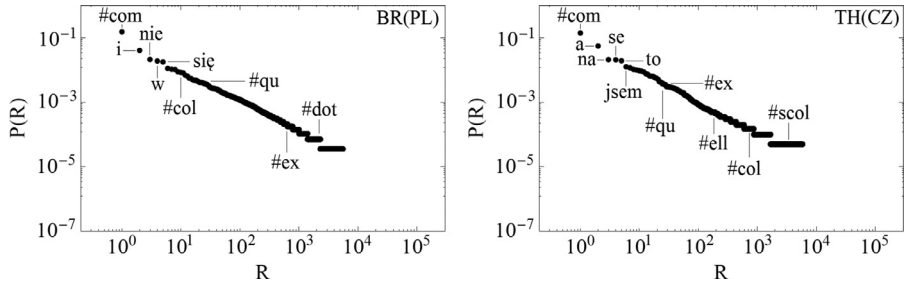

Fig. 2. The word and punctuation mark occurrence probability distributions for two texts with unusual punctuation statistics: *Bramy raju* by Jerzy Andrzejewski (Polish, left) and *Taneční hodiny pro starší pokročilé* by Bohumil Hrabal (Czech, right). The abbreviations are the same as in Fig. 1.

is the strongest for the Slavic languages (essentially $c = 0$) and the weakest, but still sizeable, for the Germanic ones. This provides a quantitative evidence that the punctuation marks included in a Zipfian plot largely restore its scaling, indeed.

3.2. Network properties for chosen words

Fig. 4 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 Table 1 the chosen network parameters are shown for the corpora.

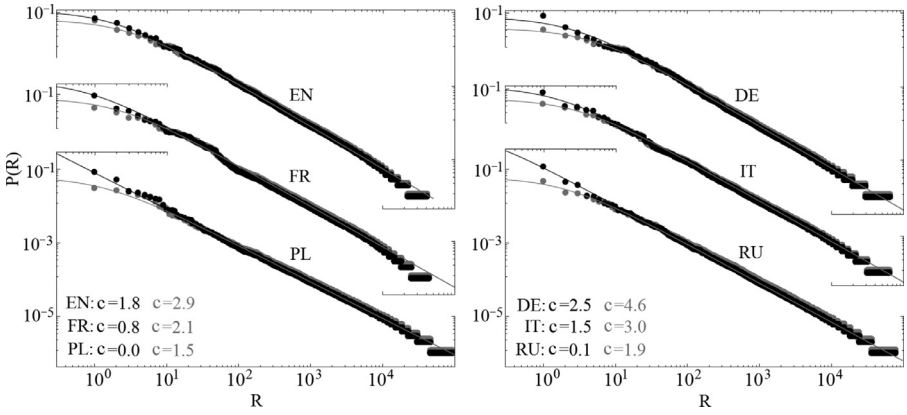

Fig. 3. The Zipf–Mandelbrot distribution fitted to the occurrence-probability distributions for corpora representing different European languages. The two following cases are distinguished: the words without the punctuation marks (grey) and the words with the punctuation marks (black). For each language and for both cases, the corresponding value of the best-fitted parameter c is given.

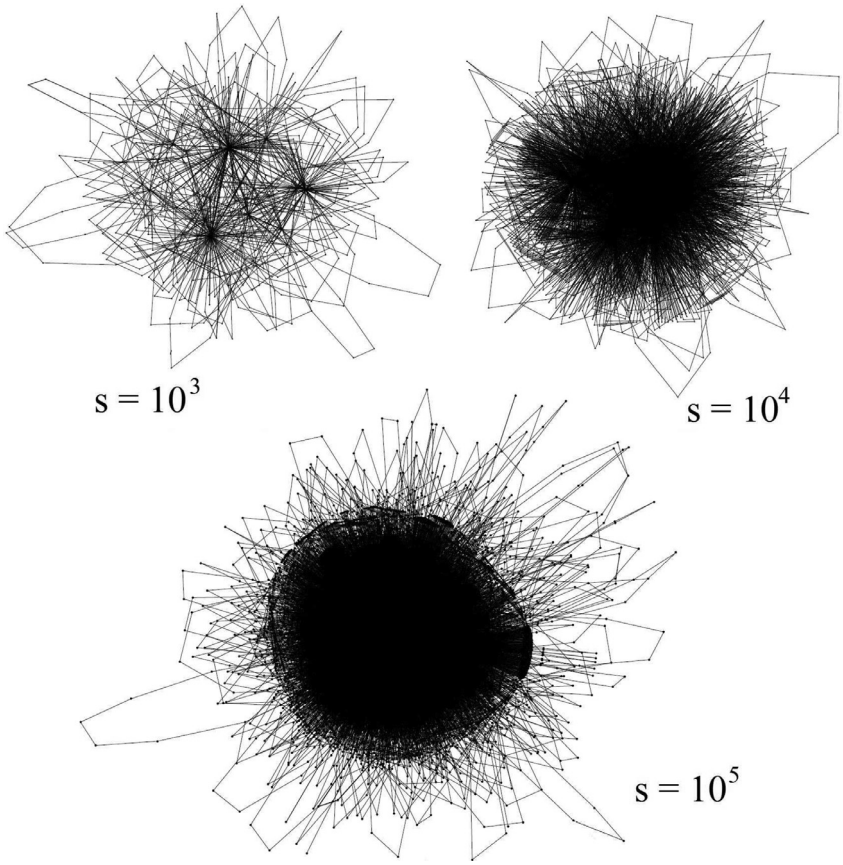

Fig. 4. Typical forms of a growing word-adjacency network created from text samples of length $s = 10^3, 10^4, 10^5$ words.

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

	English	German	French	Italian	Polish	Russian
n	40,673	65,818	44,788	60,985	89,993	91,049
e	272,501	375,452	302,243	398,796	473,611	472,133

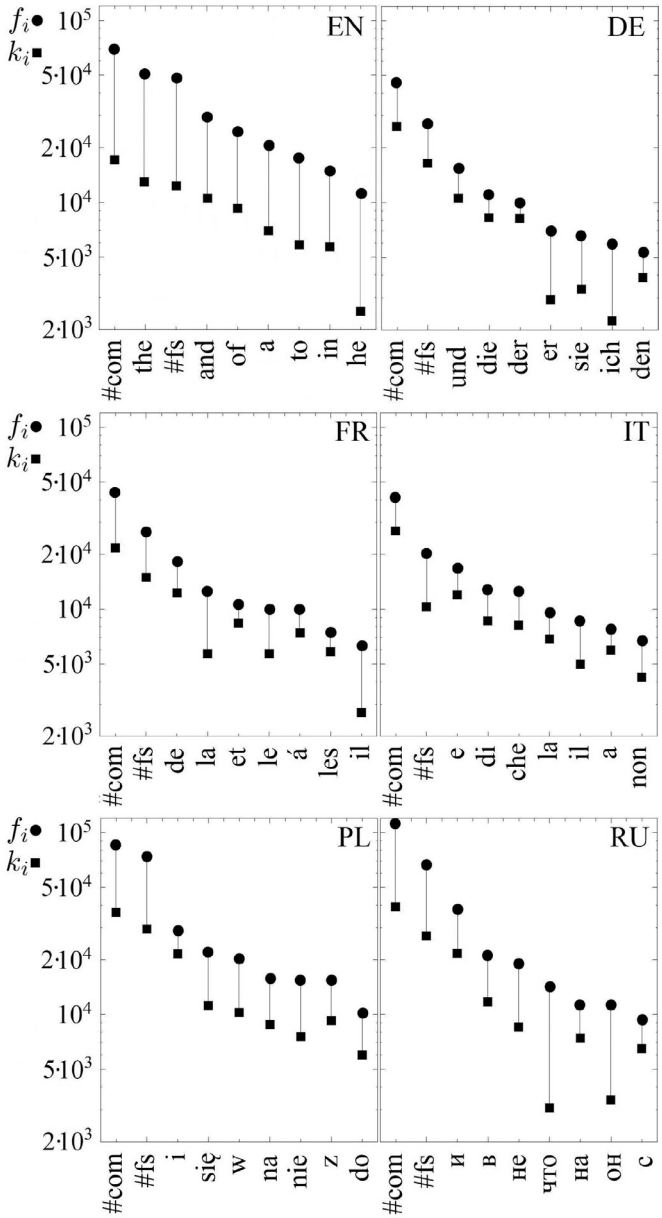

Fig. 5. Difference between the frequency f_i of the most common words, full stops, and commas (circles) and the degree k_i of the respective nodes (squares) for the Germanic (top), Romance (middle), and Slavic (bottom) languages.

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 ith 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 ith 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. 5 the difference between f_i and k_i is shown for the most frequent items in a proper order starting from the left-hand side.

In English (Fig. 5(top)), 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 $1/5 < a(i) < 1/3$. 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 the remaining five languages, more significant variability among different items is observed. In German, the pronouns: *er*, *sie*, and *ich* are represented by larger differences between both quantities ($1/6 < a(i) < 1/4$) than the punctuation marks and the other considered words ($1/3 < R < 1/2$). In French, the pronouns/articles: *le*, *la*, *il* show large differences

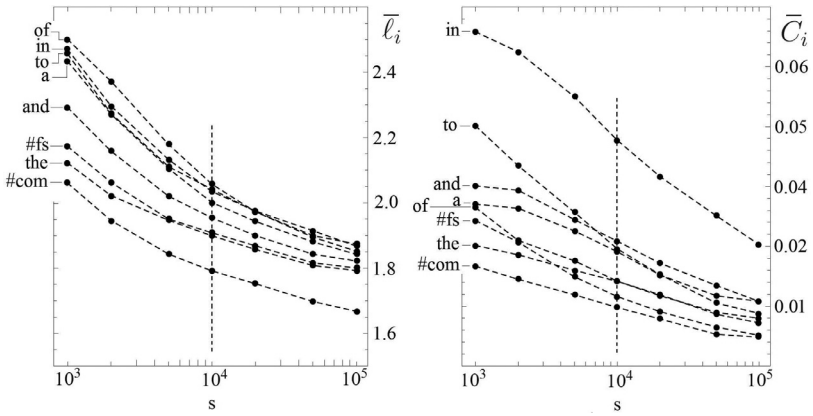

Fig. 6. The word-specific average shortest path length $\bar{\ell}_i$ (left) and the local clustering coefficient $\bar{C}_i$ (right) averaged over different text samples as functions of the sample size s for the most frequent English words, full stops (#fs), and commas (#com). The vertical line indicates the value of $s = 10^4$ used for creating Fig. 7.

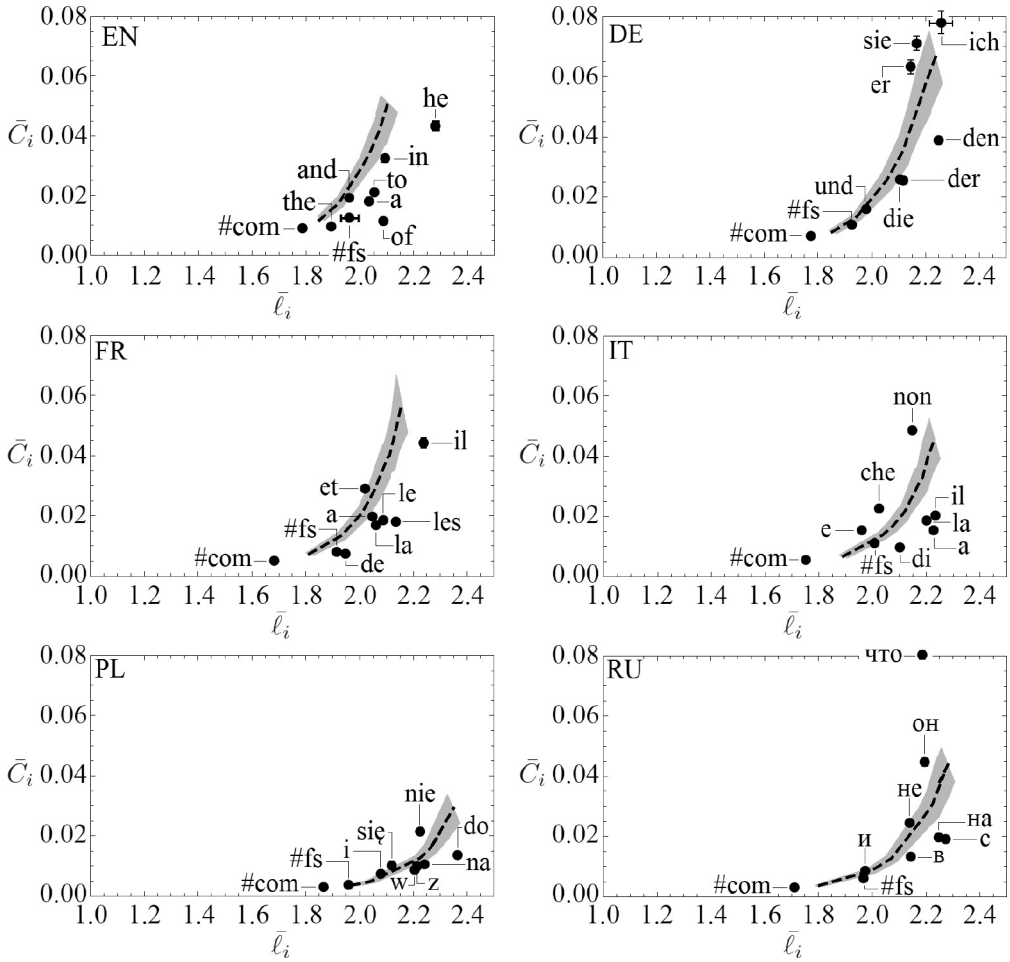

Fig. 7. Scatter plots of the word-specific average shortest-path length $\bar{\ell}_i$ and the local clustering coefficient $\bar{C}_i$ for the most frequent words, full stops #fs, and commas #com in six different European languages: English (top left), German (top right), French (middle left), Italian (middle right), Polish (bottom left), and Russian (bottom right). Error bars denote standard deviations calculated from 100 independent text samples. A null model of random word sequences (100 independent realizations) is represented by its mean (dashed line) and standard deviation (grey region).

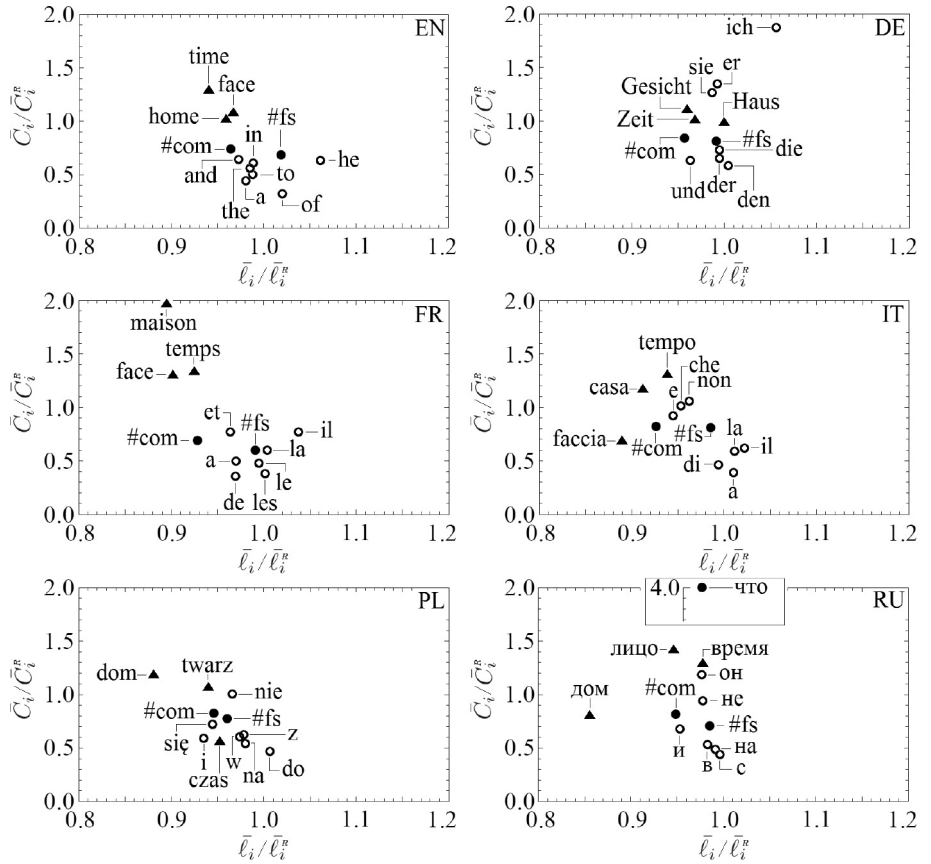

Fig. 8. Scatter plots of the word-specific average shortest-path length $\bar{\ell}_i$ and the local clustering coefficient $\bar{C}_i$ divided by their random-null-model counterparts $\bar{\ell}_i^R$ and $\bar{C}_i^R$, respectively. The same corpora as in Fig. 7 are used. In addition to the words (open circles) and the punctuation marks (full circles) used in Fig. 7, the results for three sample lexical words, semantically the same for each language, are also presented (full triangles).

Table 2
Ranks R of the lexical words used in Fig. 8.

R	English	German	French	Italian	Polish	Russian
time	75	118	145	96	228	93
face	130	185	247	527	168	124
home	264	242	157	145	429	340

up to $a(i) \approx 1/8$, the pronoun *les*, the preposition *á*, and the conjunction *et* show small differences ($a(i) \approx 3/5$), and the punctuation marks present moderate behaviour (Fig. 5(middle)). 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. 5(bottom)). The smallest difference is for a Polish conjunction *i* ($a(i) \approx 3/4$) since this word does not have any special collocation with other words. On the other hand, the punctuation marks can be collocated with specific words and this property is reflected in the largest difference between f_i and k_i ($a(i) \approx 1/3$), but nevertheless this difference does not exceed those observed for other words much. In Russian the variability between the items is also strong with the pronouns *что* and *он* exhibiting the largest differences ($1/6 < a(i) < 1/3$), while the conjunction *и* and the prepositions: *на*, *с* exhibiting the smallest ones ($a(i) \approx 3/5$). 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

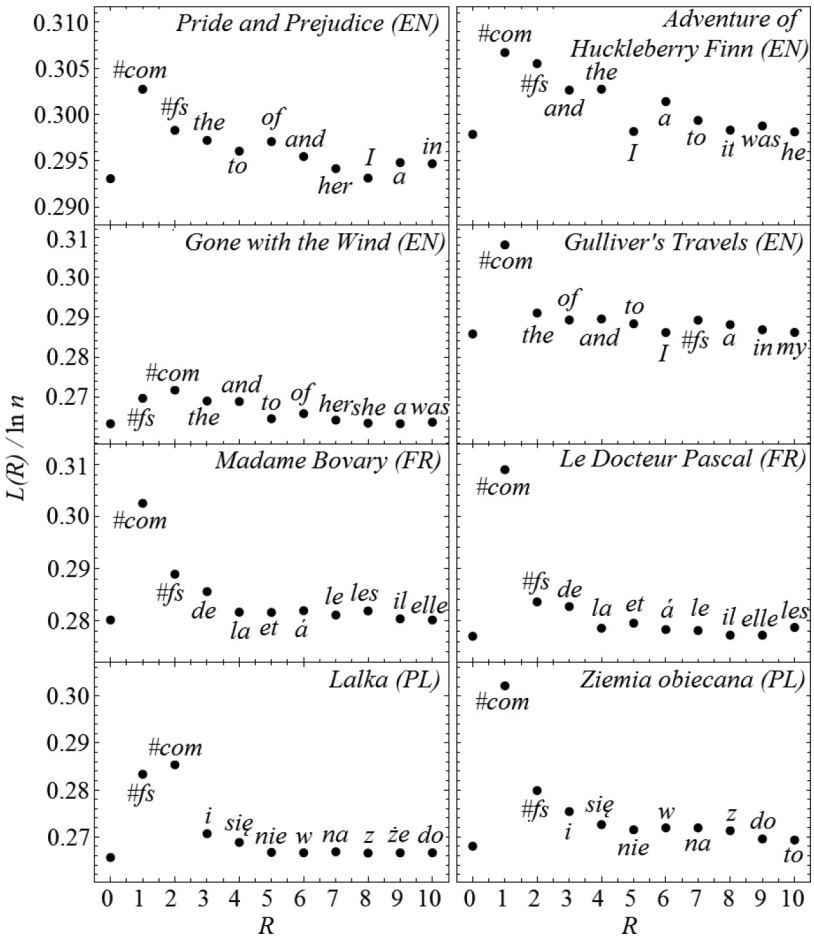

Fig. 9. The average shortest path length $L(R)$ with respect to $\ln n$ for the networks representing different text samples (novels). For each novel, the rank $R = 0$ denotes the complete network with all the n nodes, while the lower ranks $1 \leq R \leq 10$ denote the incomplete networks with $n - 1$ nodes obtained by removing a node corresponding to a word ranked R in the Zipf distribution for this novel.

other nodes in the network and it is defined as follows:

$$\ell_i = \frac{1}{n-1} \sum_{j}^{n} d(i, j), \tag{2}$$

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:

$$C_i = \frac{2e_i}{k_i(k_i - 1)}, \tag{3}$$

where e_i is the number of connections between direct neighbours of the ith node and k_i is its degree. This measure defines the density of links between direct neighbours of a given node and it can reveal membership of this node in a specific subset of strongly interconnected nodes [12]. In order to calculate ℓ_i and C_i, one has to note that both quantities depend on n [16]. This is because, according to the Heaps law, there is a non-linear dependence 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 [11]. In result the network becomes saturated gradually with increasing the sample size 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 [16]. 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 $s_c \approx 10^6$ was looped by connecting the last stop mark with the first word (this artificial link was removed from the networks, of course). 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

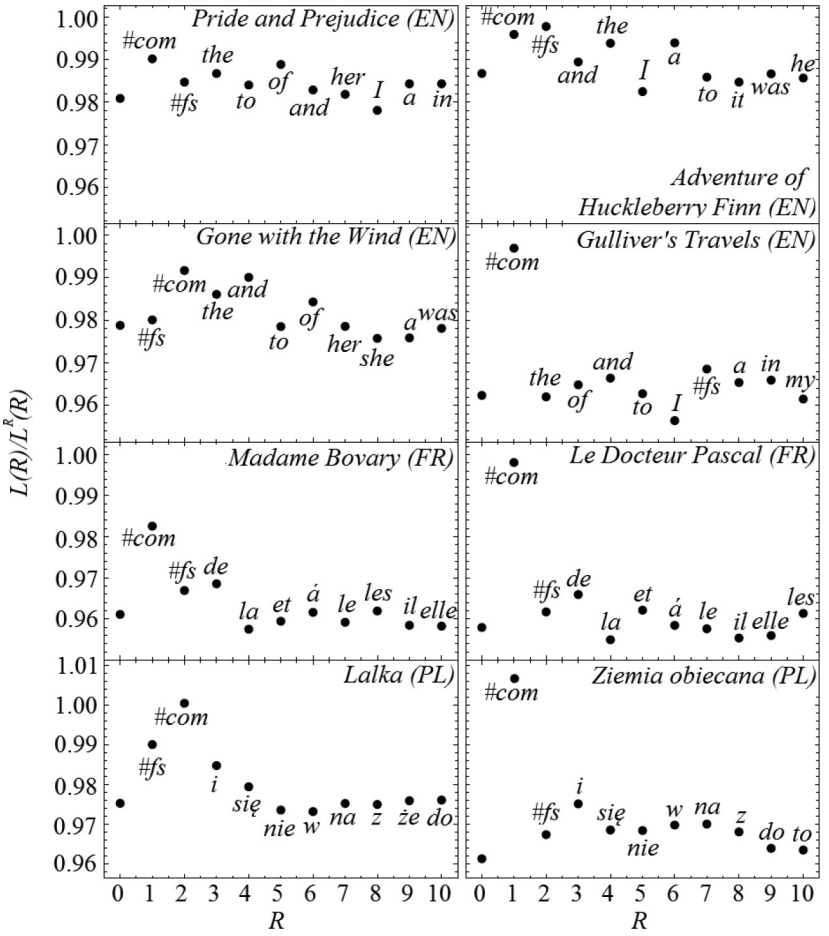

Fig. 10. The average shortest path length $L(R)$ for the networks representing different novels (the same as in Fig. 9) divided by its counterpart $L^R(R)$ calculated for the null model of random texts with the same Zipf distribution.

network parameters ℓ_i and C_i were calculated for each network realization independently for the 10 most frequent items 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 and punctuation marks is presented in Fig. 6(left) and Fig. 6(right), 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 $\bar{\ell}_i(s)$ for #fs and $\bar{C}_i(s)$ for both #fs and #com do not differ much from their counterparts representing the ordinary words, $\bar{\ell}_i(s)$ for comma is distinguished by exceptionally small values while preserving the monotonically decreasing shape of ASPL for the other objects. The results obtained for all the 6 languages are summarized in Fig. 7 in a form of scatter plots $\bar{\ell}_i$ vs. $\bar{C}_i$ for the medium sample size of $s = 10^4$. The standard errors determined for $\bar{\ell}_i$ and $\bar{C}_i$ are typically so small that they do not differ much from the symbol size in Fig. 7. The full stop and comma have rather low values of ASPL. Among the considered words, the most distinguished one is the German pronoun *ich* with a significant variability of both $\bar{\ell}_i$ and $\bar{C}_i$ among the individual sample networks. Although not explicitly shown here, the same observation refers to this word's counterparts in other languages (like *I*, *je*, *ja*, *я*), whose variability is related to particular choices of the considered texts with different narration types.

Owing to the ASPL definition, in each case the value of $\bar{\ell}_i$ is negatively correlated with the node degree k_i. LCC is also strongly anticorrelated with k_i and its empirical dependence on the node degree is roughly power-law, which agrees with the theoretical considerations for the hierarchical networks [7]. This double dependence on k_i means that $\bar{C}_i$ may also be considered a function of $\bar{\ell}_i$. We expect thus that substantial contribution to the variability of $\bar{C}_i$ and $\bar{\ell}_i$ in Fig. 7 comes from this relation. In order to show this, we calculated both quantities for the random null model, in which no correlations between the words are allowed and their occurrences are governed only by their relative frequency given by the Zipf distribution. The corpora in each language was randomly shuffled, so the constituent texts lost their significance as they became just meaningless word sequences. Then we constructed the corresponding word-adjacency networks and calculated both ASPLs and LCCs for the nodes representing the same words as in Fig. 7. We repeated this procedure 100 times inde-

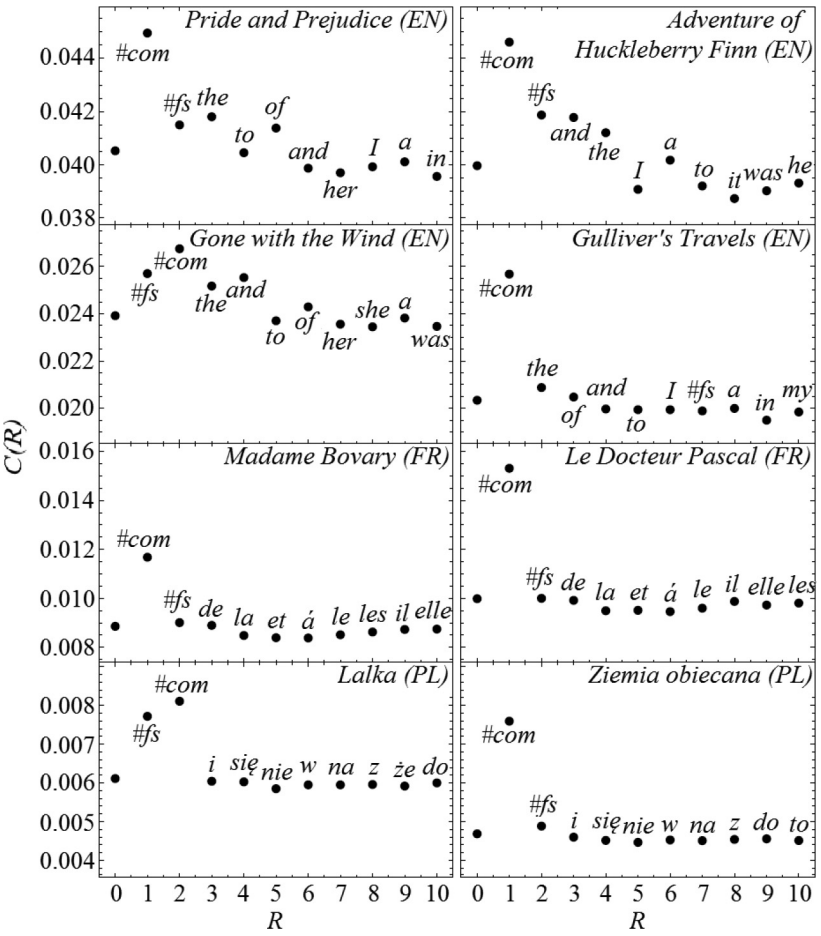

Fig. 11. The global clustering coefficient $C(R)$ for the networks representing different novels. For each novel, the rank $R = 0$ denotes the complete network with all n nodes, while the lower ranks $1 \leq R \leq 10$ denote the incomplete networks with $n - 1$ nodes obtained by removing a node corresponding to the word ranked R in the Zipf distribution for this novel.

pendently. Indeed, for each language we obtained an approximately power-law relation: $\bar{C}_i(\bar{\ell}_i) \sim \bar{\ell}_i^{\gamma}$ with $\gamma > 0$ (denoted by a dashed line in Fig. 7).

The points that denote the $(\bar{\ell}_i, \bar{C}_i)$ coordinates for the particular items in Fig. 7 are distributed along this functional dependence. This means that the item positions on the scatter plots are strongly influenced by these items' frequencies, while the actual grammar- and context-related contributions to $\bar{C}_i$ and $\bar{\ell}_i$ are less evident. Therefore, we decided to remove the frequency-based contributions by dividing the empirical values by their average random-model counterparts: $\bar{C}_i^{\mathrm{R}}$ and $\bar{\ell}_i^{\mathrm{R}}$. The resulting positions of the items are shown in Fig. 8. Now it is more evident than in Fig. 7 that both the high-frequency words and the punctuation marks occupy similar positions and no quantitative difference can be identified that is able to distinguish between both groups. In this figure, we also show these quantities calculated for three sample words chosen randomly from more distant parts of the Zipf plot: *time* ($R = 75$ in the English corpus), *face* ($R = 130$), and *home* ($R = 264$), as well as their semantical counterparts in the other languages (occupying different ranks, see Table 2). Obviously, each of these words may also have other, non-equivalent meanings in distinct languages and, while some languages use inflection, the other ones do not, which inevitably contribute to the rank differences. In contrast to the most frequent words discussed before, these words are significantly less frequent, which can itself lead to some differences in the statistical properties as compared to the top-ranked words. Therefore, they are not shown in Fig. 7, because their local clustering coefficient significantly exceeds the vertical axis upper limit. In Fig. 8, the sample lexical words are located in different places for different languages, but typically their average position is more or less shifted towards the upper left corner of the plots. This effect is the most pronounced for French, then for English, Russian, Italian, and Polish, while it is absent for German. This visible shift may originate from either the statistical fluctuations among the words, the statistical fluctuations among the texts selected for the corpora, or be a genuine effect for the less frequent words, the parts of speech, and/or a general property of the lexical words in specific languages. However, since our sample of the medium-ranked words is small, at present we prefer not to infer any decisive conclusions from this result as we plan to carry out a related, comprehensive

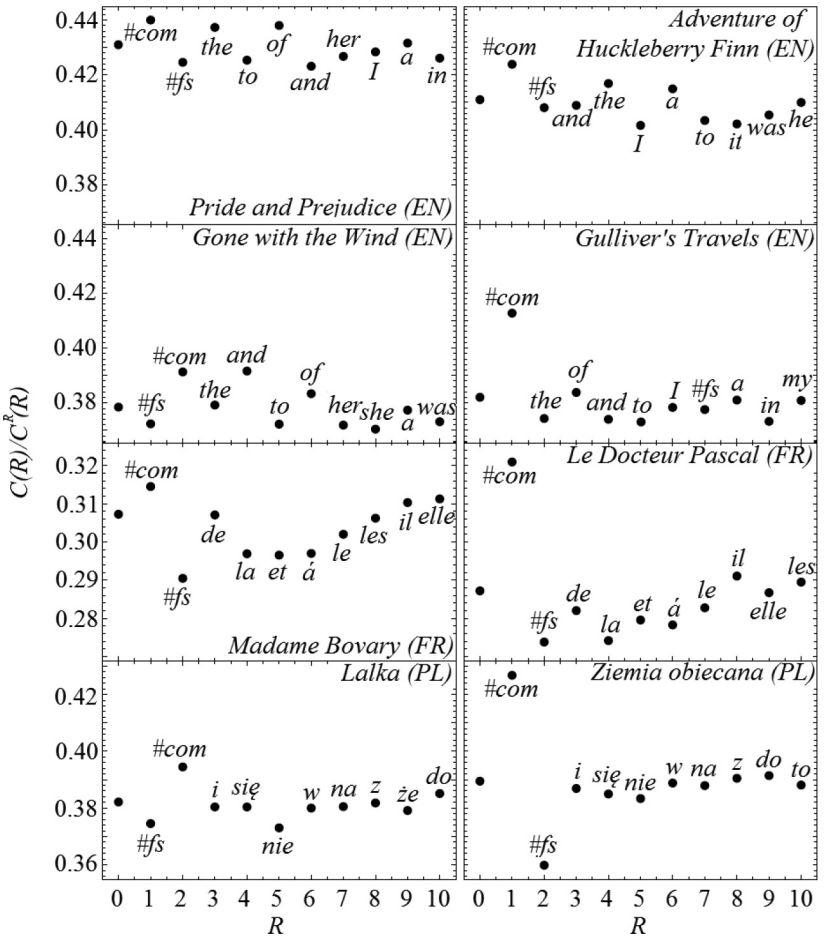

Fig. 12. The global clustering coefficient $C(R)$ for the networks representing different novels (the same as in Fig. 11) divided by its counterpart $C^{R}(R)$ calculated for the null model of random texts with the same Zipf distribution.

study in near future. Nevertheless, we stress here that such displacements exhibited in Fig. 8 by the words of medium frequency by no means contradict our main statement that the punctuation marks show similar statistical properties as the most frequent words.

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:

$$r = \frac{\sum_{ij}(\delta_{ij} - \frac{k_i k_j}{2e})}{\sum_{ij}(k_i\delta_{ij} - \frac{k_i k_j}{2e})}, \tag{4}$$

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. Due to the same reason as before, we first calculate the corresponding quantities L^{R}, C^{R}, and r^{R} for the randomized text samples (100 independent realizations) and consider them the reference values determined solely by the frequencies of particular items and by neither grammar nor context. We thus consider the values of $L(R)/L^{R}(R)$ (Fig. 9), $C(R)/C^{R}(R)$ (Fig. 11), and $r(R)/r^{R}(R)$ (Fig. 13) and expect them to be related to grammar and context largely. For different text samples (novels), we compare the corresponding values calculated for a complete network with all the nodes present (denoted by the abscissa $R = 0$) and for 10 incomplete networks obtained by removing a given highly ranked node according to the Zipf distribution ($1 \le R \le 10$).

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 - see Fig. 9, but a rule is that, statistically, the lower the rank (the larger R) is, the smaller is the change in $L(R)$ (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. This means

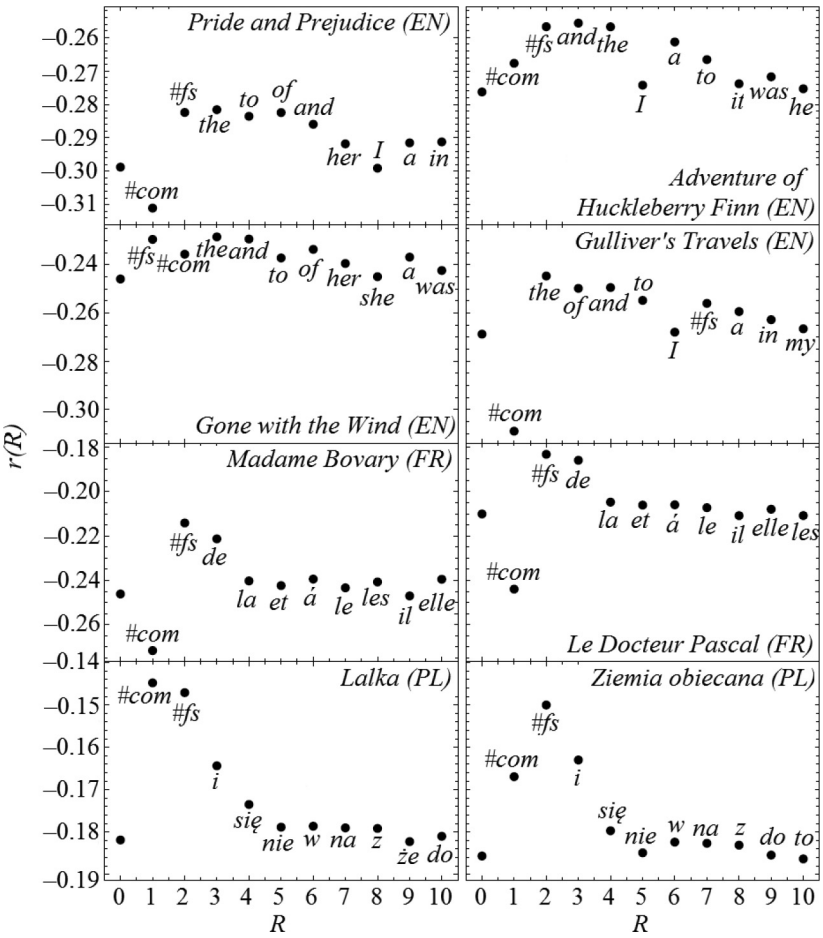

Fig. 13. The global assortativity index $r(R)$ for the networks representing different novels. For each novel, the rank $R = 0$ denotes the complete network with all n nodes, while the lower ranks $1 \leq R \leq 10$ denote the incomplete networks with $n - 1$ nodes obtained by removing a node corresponding to the word ranked R in the Zipf distribution for this novel.

that in a typical situation $L(R)$ alters its value the most for comma and the full stop since they occupy the highest ranks, while the observed changes for the function words are smaller. This picture substantially changes if we look at the rescaled quantity: $\lambda(R) = L(R)/L^{\mathrm{R}}(R)$ that is free of an item's frequency contribution to ASPL. Fig. 10 shows that, except for #com, $\lambda(R)$ does not exhibit any significant dependence on R and excluding a particular node does not alter its value much as compared to the complete network. Typically, the rescaled ASPL is restricted to a narrow range of $0.95 < \lambda(R) < 1.0$ and this means that the correlations present in the text samples shorten effectively the paths between the nodes as compared to the random network, but this is a small effect. The case of comma is slightly different as, for some texts, the network without this node shows $\lambda \approx 1.0$, i.e., the residual network has the same $L(R)$ as the random one. Presence of this property of comma is text-dependent and it does not seem to be a property of written language. Moreover, it should be stressed that, even if one considers comma, the range of the $\lambda(R)$ variability for different nodes is small.

The global clustering coefficient $C(R)$ and the assortativity index $r(R)$ present a more variable behaviour after removing a hub as these quantities can either increase, remain stable, or decrease. This behaviour obviously depends on a contribution of each particular node to C and r for $R = 0$. For the clustering coefficient, a statistical rule is that without particular nodes $C(R)$ does not differ much from its complete-network counterparts. Only for the node representing comma, $C(R)$ can increase more significantly and the network becomes more clustered (Fig. 11). This can partially be explained by an observation that in all the considered languages commas can mediate words whose direct neighbourhood is unlikely due to rules of grammar. Since $C(R)$ depends on an item's frequency, in Fig. 12 we show the rescaled coefficient: $\kappa(R) = C(R)/C^{\mathrm{R}}(R)$ whose values are related to the random model. Now the networks without #com and the ones without other nodes show comparable values of κ with only small difference (up to 10%) for some texts.

As regards the assortativity index $r(R)$, the majority of hubs in the word-adjacency networks (like, e.g., #fs, articles, and the most frequent conjunctions) can be considered disassortative separators, so after their removal, the overall assortativity index increases (Fig. 13). Of course, since this is only a statistical observation, particular cases may show different behaviour

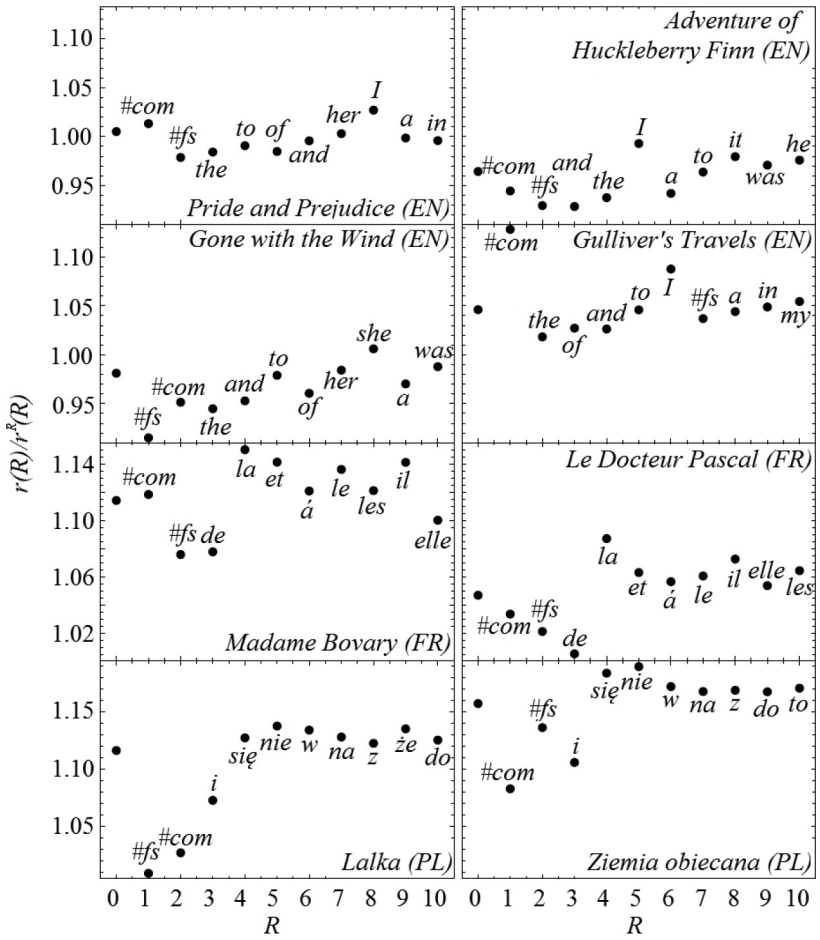

Fig. 14. The global assortativity index $r(R)$ for the networks representing different novels divided by its counterpart $r^{R}(R)$ calculated for the null model of random texts with the same Zipf distribution.

like, e.g., comma, which sometimes acts like a disassortative separator and sometimes like an assortative one. In order to remove the approximately monotonous dependence of r on rank R, we look at the rescaled assortativity index: $\rho(R) = r(R)/r^{R}(R)$. We see that now this dependence is absent and that both the punctuation marks and the function words exhibit comparable values of ρ (see the narrow range of the vertical axis in each panel of Fig. 14).

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 the 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 the most common punctuation marks and words using two approaches. We observed that the punctuation marks locate themselves exactly on or in a close vicinity of the power-law Zipfian regime as if they were ordinary words. Moreover, their inclusion acts towards restoring of the Zipf power-law from the more flat Zipf–Mandelbrot behaviour. 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, the punctuation marks are more important than typical nodes: they play a role of the 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 and medium common words (both the function and the lexical ones) if one investigates their statistical properties. Since the punctuation marks have also non-neglectable meaning, we advocate their inclusion in any type of the word-occurrence and the word-adjacency analysis making it to be more complete. Incorporation of the punctuation marks into an analysis extends its dimensionality and, therefore, it opens more space for possible manifestation of some previously unobserved effects. That this can in fact be fruitful and bring important results, the best example is Ref. [8] where we showed that the sentence length variability can be multifractal for specific (written with the stream-of-consciousness narration) group of texts, while for other texts it remains monofractal. Multifractality is inherently accompanied by burstiness. In the present context this burstiness in the sentence length thus translates itself into analogous effects in the recurrence times (measured by a separation of two consecutive occurrences of the same item) of the full stops. At the same time, however, the recurrence times of the most frequent words appear to be much less bursty as it was also documented in the same Ref. [8]. This latter effect goes in parallel with an observation made earlier [1] for the most frequent words. Interestingly, according to the same paper [1], the burstiness is however often observed for the non-function words of high and medium ranks. The related intricacy in fact further supports our thesis that, from the statistical point of view, the punctuation marks are surprisingly similar to the regular words, even though at some angles they resemble more the most frequent function words, while at other angles they resemble more the non-function words. We expect more interesting results will be obtained in future from analyses, in which the punctuation marks are not neglected.

Acknowledgement

We thank the anonymous referees for very interesting and insightful suggestions that led to significant extensions and improvement of this paper.

Appendix

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

English: George Orwell *1984**, Mark Twain *Adventures of Huckleberry Finn*, Herman Melville *Moby Dick**, 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*, Jerzy Andrzejewski *Bramy raju*.

Russian: Lev Tolstoy *Анна Каренина* (*Anna Karenina*)*, *Война и мир* (*Voyna i mir*)*, *Воскресение* (*Voskreseniye*)*, *Fyodor Dostoyevsky Бесы (Besy)**, *Братья Карамазовы (Brat'ya Karamazovy)**.

Czech: Bohumil Hrabal Taneční hodiny pro starší a pokročilé.

References

[1] E.G. Altmann, J.B. Pierrehumbert, A.E. Motter, Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words, PLoS ONE 4 (2009) e7678.
[2] 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, Physica A 391 (2012) 4406–4419.
[3] D.R. Amancio, A complex network approach to stylometry, PLoS ONE 10 (2015) e0136076.
[4] P.W. Anderson, More is different, Science 177 (1972) 393–396.
[5] M. Ausloos, Punctuation effects in english and esperanto texts, Physica A 389 (2010) 2835–2840.
[6] S.N. Dorogovtsev, J.F.F. Mendes, Language as an evolving word web, Proc. R. Soc. Lond. B 268 (2001) 2603–2606.
[7] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, Pseudofractal scale-free web, Phys. Rev. E 65 (2002) 066122.
[8] S. Drożdż, P. Oświęcimka, A. Kulig, J. Kwapień, 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.
[9] J.-B. Estoup, Gammes sténographiques, Methodes et exercises pour l'acquisition de la vitesse, Institut Sténographique de France, 1916.
[10] R. Ferrer-i Cancho, R.V. Solé, The small world of human language, Proc. R. Soc. Lond. B: Biol. Sci. 268 (2001) 2261–2265.
[11] M. Gerlach, E.G. Altmann, Stochastic model for the vocabulary growth in natural languages, Phys. Rev. X 3 (2013) 021006.
[12] I. Grabska-Gradzińska, A. Kulig, J. Kwapień, S. Drożdż, Complex network analysis of literary and scientific texts, Int. J. Mod. Phys. C 23 (2012) 1250051.
[13] H.S. Heaps, Information Retrieval: Computational and Theoretical Aspects, Academic Press, Orlando, 1978.
[14] G. Herdan, Type-token Mathematics, A Textbook of Mathematical Linguistics, Mouton, 's-Gravenhage, 1960.
[15] A. Kao, S.R. Poteet, Natural Language Processing and Text Mining, Springer Science & Business Media, Berlin, 2007.
[16] A. Kulig, S. Drożdż, J. Kwapień, P. Oświęcimka, Modeling the average shortest-path length in growth of word-adjacency networks, Phys. Rev. E 91 (2015) 032810.
[17] J. Kwapień, S. Drożdż, A. Orczyk, Linguistic complexity: english vs. polish, text vs. corpus, Acta Phys. Pol. A 117 (2010) 716–720.
[18] J. Kwapień, S. Drożdż, Physical approach to complex systems, Phys. Rep. 515 (2012) 115–226.
[19] H. Liu, Statistical properties of chinese semantic networks, Chin. Sci. Bull. 54 (2009) 2781–2785.
[20] B.B. Mandelbrot, An information theory of the statistical structure of language, in: W. Jackson (Ed.), Communication Theory, Academic Press, New York, 1953, pp. 503–512.

[21] B. Mandelbrot, Information theory and psycholinguistics: a theory of words frequencies, in: P. Lazafeld, N. Henry (Eds.), Readings in Mathematical Social Science, MIT Press, Cambridge, 1966.
[22] M. Markosova, Network model of human language, Phys. A 387 (2008) 661–666.
[23] A.P. Masucci, G.J. Rodgers, Network properties of written human language, Phys. Rev. E 74 (2006) 026102.
[24] M.A. Montemurro, Beyond the Zipf-Mandelbrot law in quantitative linguistics, Physica A 300 (2001) 567–578.
[25] W. Piotrowska, X. Piotrowska, Statistical parameters in pathological text, J. Quant. Ling. 11 (2004) 133–140.
[26] The project gutenberg website, www.gutenberg.org.
[27] G.K. Zipf, Selective Studies and the Principle of Relative Frequency in Language, MIT Press, Cambridge, 1932.
[28] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, Cambridge, 1949.