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/© 2016 Elsevier Inc. All rights reserved.

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