Taking a breath when reading: on the distribution of punctuation in written language

Tomasz Stanisz

Complex Systems Theory Department, IFJ PAN, Kraków

December 14, 2023

Natural language and complexity

Complex systems:

- multilevel, hierarchical structure
- complicated interactions between system's constituents
- global properties of the system are not straightforwardly derived solely from the properties of the system's individual elements
  $\rightarrow$ emergence, sometimes considered as defining property of complexity

"more is different"

"the whole is something beside the parts"

Natural language and complexity

An example of a characteristic of natural language related to complexity – the importance of context, making lingustic structures acquire meanings or roles different from the original ones.

Examples of effects/features typical for complex systems, which are observed in natural language:

- power laws
- long-range correlations – and resulting fractal and multifractal structures
- complicated (*complex*) organization in network representation

Various perspectives on natural language

Language can be understood as

- a set of symbols and rules
- an organism's ability to generate sounds
- a communication tool
- a logical system of notions guiding the thinking process
- a social and cultural phenomenon
- ...

$\rightarrow$ interdisciplinary approach, utilizing tools from various scientific disciplines

Quantitative study of language – attempts to identify statistical laws describing language and aims to explain their origin.

Example of a linguistic law: Zipf's law

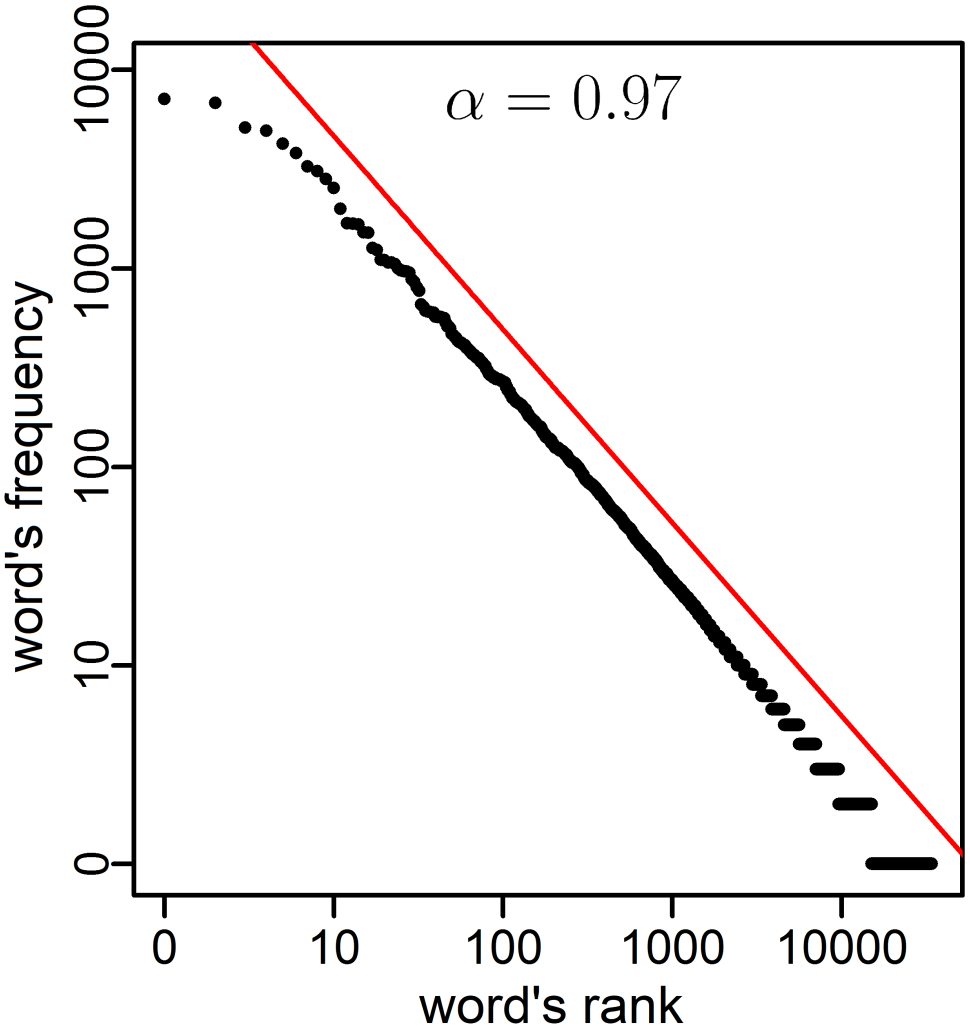
For a (sufficiently long) text:

- count the frequency ω (the number of occurrences) of each word in the text;
- rank the words by their decreasing frequency (the most frequent word gets the rank $R = 1$, the second most frequent word gets the rank $R = 2$, etc.);
- Zipf's law states that

$$\omega(R) \propto \frac{1}{R^{\alpha}} \quad \text{with } \alpha \approx 1.$$

Zipf's law is universally observed across languages – including extinct and artificial languages.

Rank-frequency distribution in *Lalka* by B. Prus



Example of a linguistic law: Zipf's law

Zipf's original explanation of the observed relationship – *principle of least effort*:

- Assume that the cost E_R of using a word with rank R is proportional to the word's length (number of letters); the average cost per word is

$$\langle E \rangle = \sum p_R E_R,$$

  where p_R is the probability of using a word with rank R.
- The "average amount of information" per word is expressed by information entropy:

$$H = -\sum p_R \log_2 p_R,$$

- The principle of least effort states that word frequency distribution in language is such that the transmission of information is cost-efficient, that is, it minimizes the quantity $\langle E \rangle / H$. This minimization gives the distribution described by Zipf's law.

It is an influential idea, but there are alternative explanations to Zipf's law; so far there is no single universally-accepted explanation.

The distribution of punctuation

Problem:
What can be said, from a quantitative perspective, about punctuation usage in written language?
– a general question, but not studied before

Motivation:

- Punctuation marks $\approx$ "breaks" in texts, dividing texts into pieces and introducing certain kind of organization.

- The arrangement of punctuation marks in texts is determined by
  – grammar,
  – logic,
  – human's capability of language processing (it is difficult to read and comprehend very long sequences of words without any breaks).

- Physical effect of a punctuation mark – making a pause or taking a breath when reading aloud.

The distribution of punctuation

Approaching the problem of punctuation distribution:

- introducing punctuation into the text $\leftrightarrow$ a random process in which after each consecutive word written, the author decides randomly whether to put a punctuation mark or not
- no distinction between punctuation marks is made – any punctuation mark from the list . ? ! ... , – ; : ( ) is just a "break"
- each decision $\leftrightarrow$ a Bernoulli trial (putting a punctuation mark is a "success", and not putting a punctuation mark is a "failure")
- writing a piece of text between a punctuation mark and the next punctuation mark $\leftrightarrow$ performing consecutive trials until the first success
- the distribution of the distances between consecutive punctuation marks (measured by the number of words) $\leftrightarrow$ the distribution of the number of trials required to get the first success
(*assuming that the process behaves in the same way over the whole text*)

The distribution of punctuation

- the simplest idea: consecutive trials are independent, probability of success is constant and equal to p
- this makes the distribution of the distances between consecutive punctuation marks to be the geometric distribution
- however, geometric distribution seems not to be representing empirical data well $\rightarrow$ more general distribution is needed
- a possible generalization of geometric distribution: the discrete Weibull distribution
- generalization consequence: consecutive trials are not independent, the probability of success changes with the number of already performed unsuccessful trials

Discrete Weibull distribution

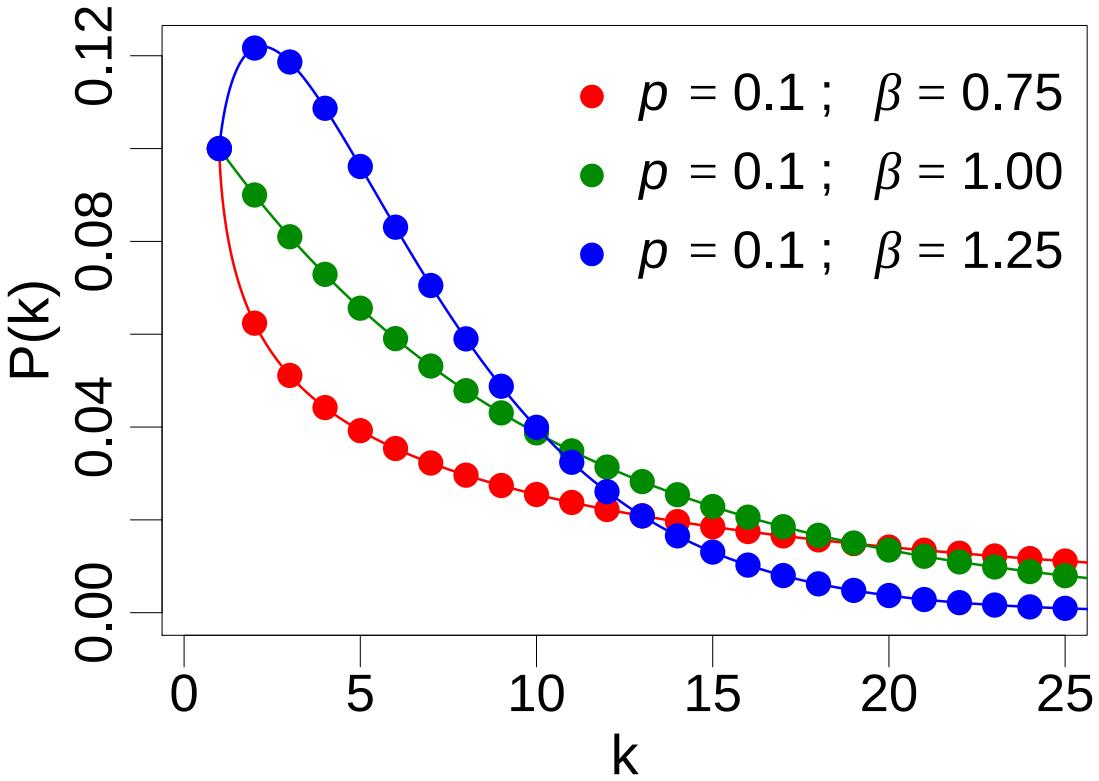
Cumulative distribution function – geometric distribution:

$$F(k) = 1 - (1-p)^k$$

Cumulative distribution function – discrete Weibull distribution:

$$F(k) = 1 - (1-p)^{k^\beta}$$

Discrete Weibull distribution – probability mass function:



Interpreting parameters of the discrete Weibull distribution
$\rightarrow$ using hazard function

Hazard functions

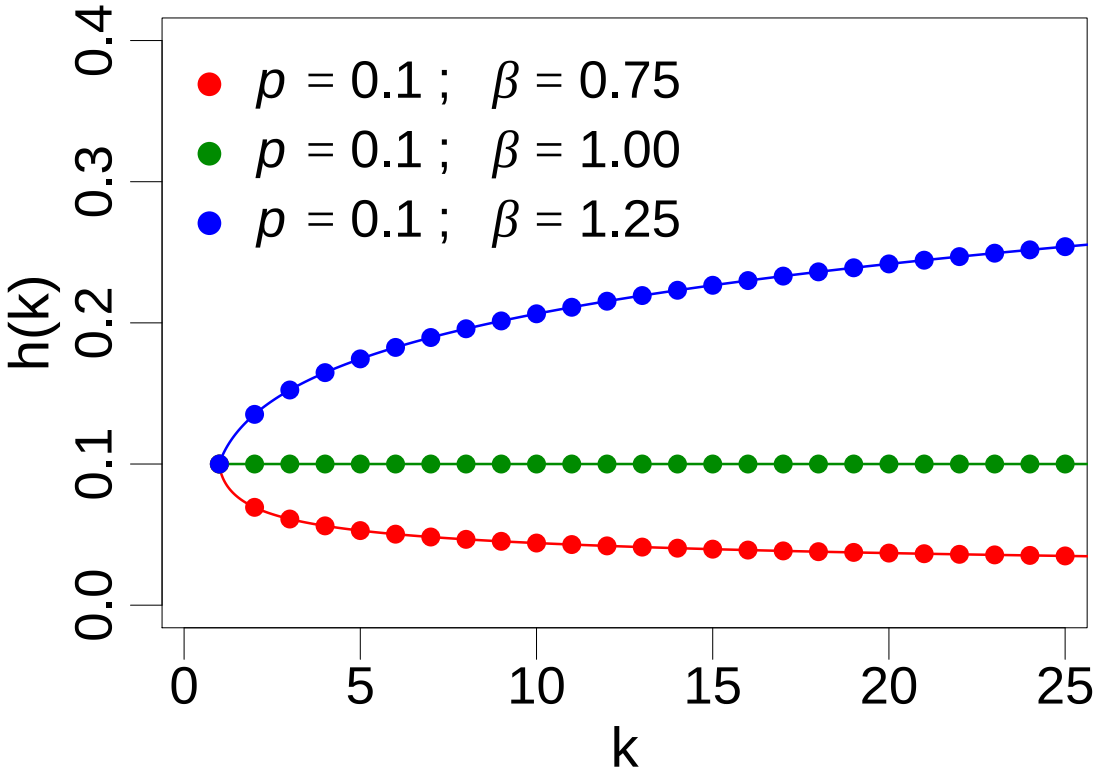
Hazard function h – a function describing the probability of obtaining the **first** success at a given trial

$h(k) =$ the conditional probability that a success occurs on the k-th trial, given that it has not occurred in the preceding $k-1$ trials

Hazard function for the discrete Weibull distribution:

$$h(k) = 1 - (1-p)^{k^{\beta} - (k-1)^{\beta}}$$

Discrete Weibull distribution – hazard functions:



Interpreting distribution parameters:

- $p = h(1)$ is the probability of placing a punctuation mark right after the first word since the last punctuation mark.
- β influences how rapidly the "pressure" on placing a punctuation mark changes with the growing number of words since the last punctuation mark.

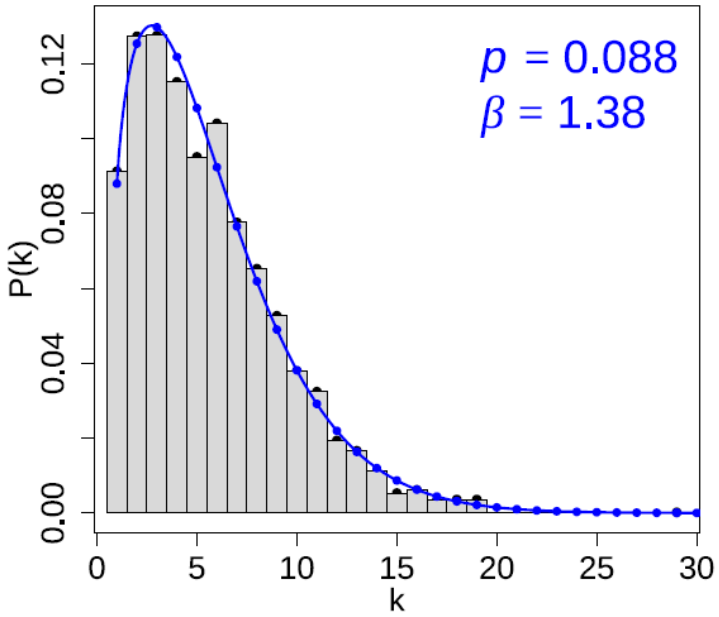
Checking the agreement of the discrete Weibull distribution with the data (empirical distances between consecutive punctuation marks) – the proposed distribution correctly approximates empirical distribution

histograms – empirical data; blue dots – fitted discrete Weibull distributions

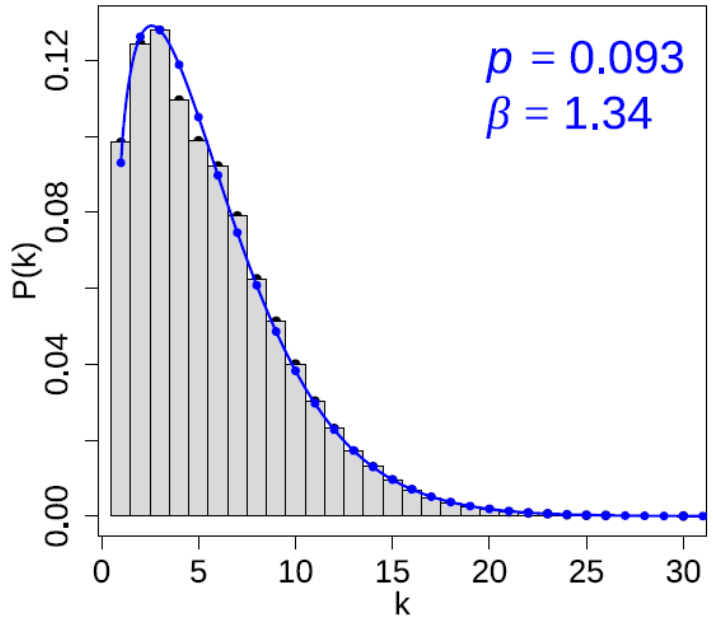
similar results are obtained for 240 texts in 7 European languages (English, German, French, Italian, Spanish, Polish, Russian)

typical values of the parameters: 0.05–0.20 for p, and 1.0–1.7 for β

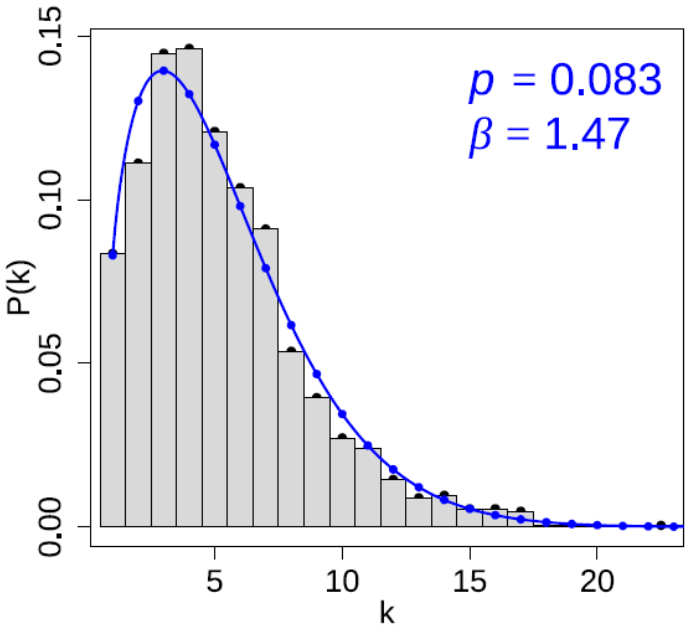
Alice's Adventures in Wonderland - L. Carroll



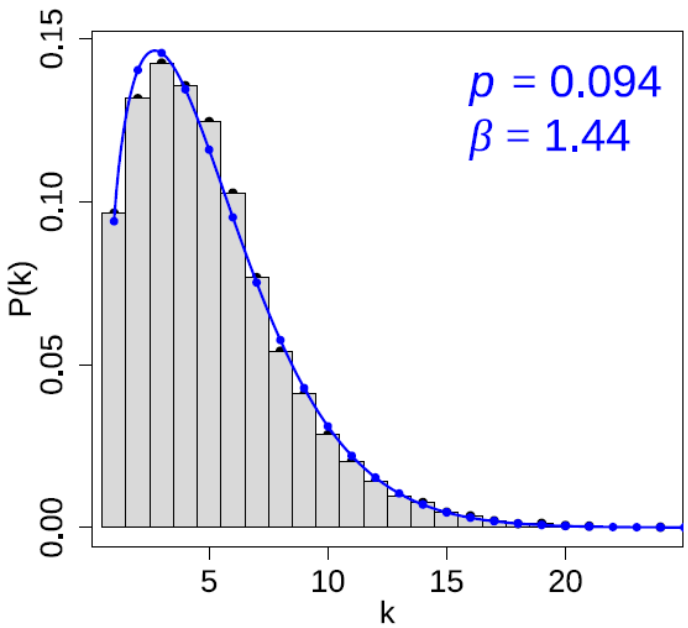
David Copperfield - C. Dickens



Le Petit Prince - A. de Saint-Exupéry



Quo vadis - H. Sienkiewicz



Sentences vs word sequences between punctuation marks

- The distances between consecutive punctuation marks seem to be universally described by the discrete Weibull distribution.
- The same is not observed for sentence lengths – there is no general rule applying to sentences (no single distribution that would correctly fit to empirical data).
- Hence, it can be stated that distances between consecutive punctuation marks behave more regularly than sentence lengths.
- This is also confirmed by certain results from time series analysis (in time series representing sentence lengths, the parameters characterizing correlations have wider range of variability that in series representing the distances between consecutive punctuation marks)
- Conclusion: the partition of a text into pieces determined by consecutive punctuation marks (of any type) is more "fundamental" than the partition into sentences (subject to more rigor).
- Results are in agreement with the common perception that punctuation marks are to a certain extent interchangeable.

Punctuation marks' interchangeability

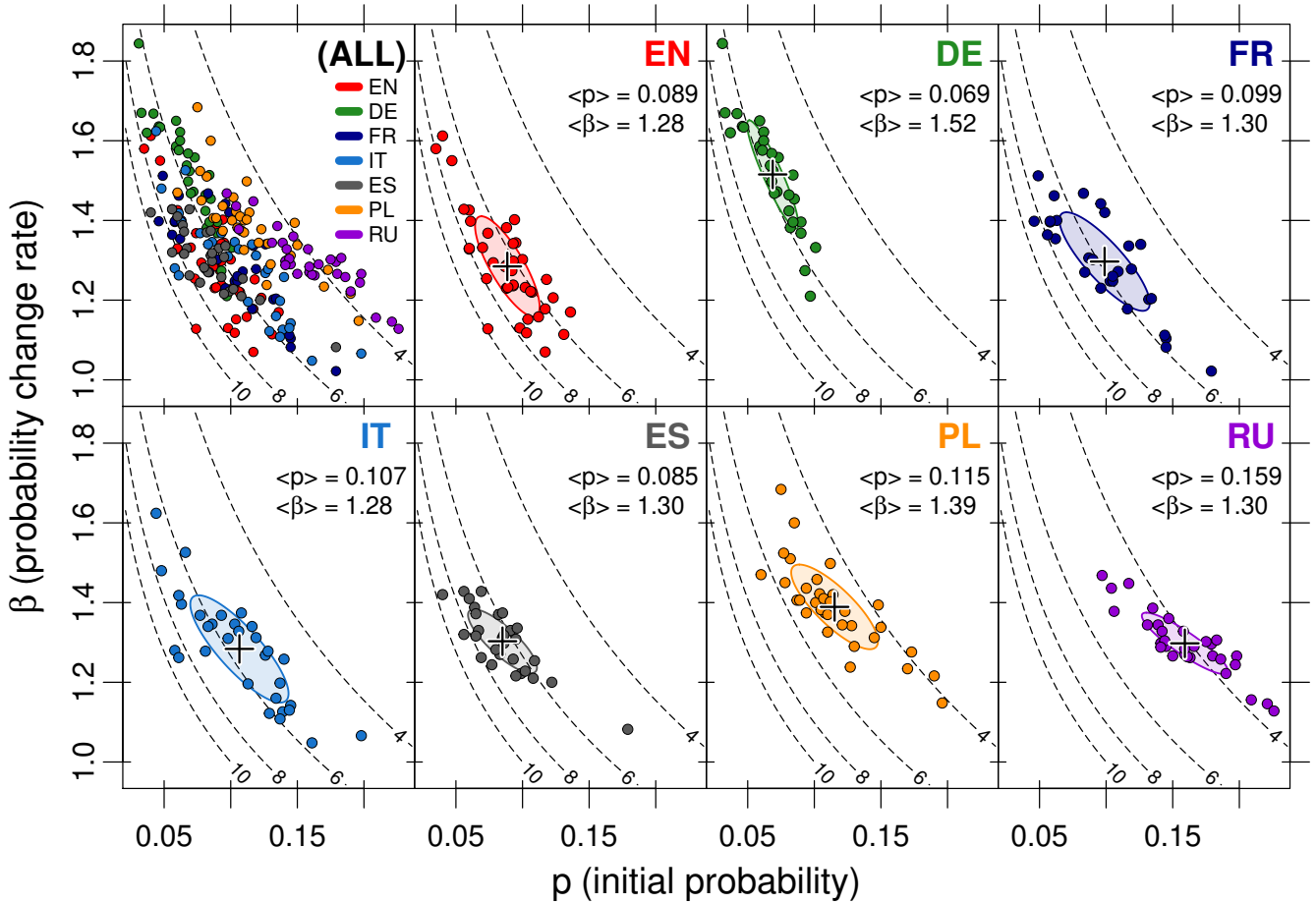
Like many other complex systems, natural language is studied from a variety of perspectives and attracts diverse academic disciplines, ranging from humanities to formal and natural sciences. One of the directions of research focuses on language's quantitative properties; it aims at identifying statistical laws characterizing various elements of language and their mutual relations – a famous example of such laws is Zipf's law, describing the distribution of word frequencies in texts. An interesting and yet unexplored issue is the question about the statistical properties of punctuation, which is responsible for introducing a specific organization into written language – punctuation marks divide texts into logically and grammatically coherent parts, clarify the meaning of potentially ambiguous phrases, and indicate when to take a breath when reading aloud. It turns out that certain features of punctuation seem to be largely universal across languages – for example, its distribution can be characterized by just two parameters which can be quite easily interpreted. On the other hand, the values of these parameters for texts in different languages might differ significantly and indicate features specific to particular languages.

Like many other complex systems, natural language is studied from a variety of perspectives and attracts diverse academic disciplines – ranging from humanities to formal and natural sciences. One of the directions of research focuses on language's quantitative properties. It aims at identifying statistical laws characterizing various elements of language and their mutual relations. A famous example of such laws is Zipf's law, describing the distribution of word frequencies in texts. An interesting and yet unexplored issue is the question about the statistical properties of punctuation, which is responsible for introducing a specific organization into written language. Punctuation marks divide texts into logically and grammatically coherent parts, clarify the meaning of potentially ambiguous phrases, and indicate when to take a breath when reading aloud. It turns out that certain features of punctuation seem to be largely universal across languages; for example, its distribution can be characterized by just two parameters which can be quite easily interpreted. On the other hand, the values of these parameters for texts in different languages might differ significantly and indicate features specific to particular languages.

- the only difference - punctuation marks (and capital letters, if needed)
- both texts have the same meaning and are grammatically valid
- using *some* punctuation mark (a "break") more important than using a specific punctuation mark

Differences between languages

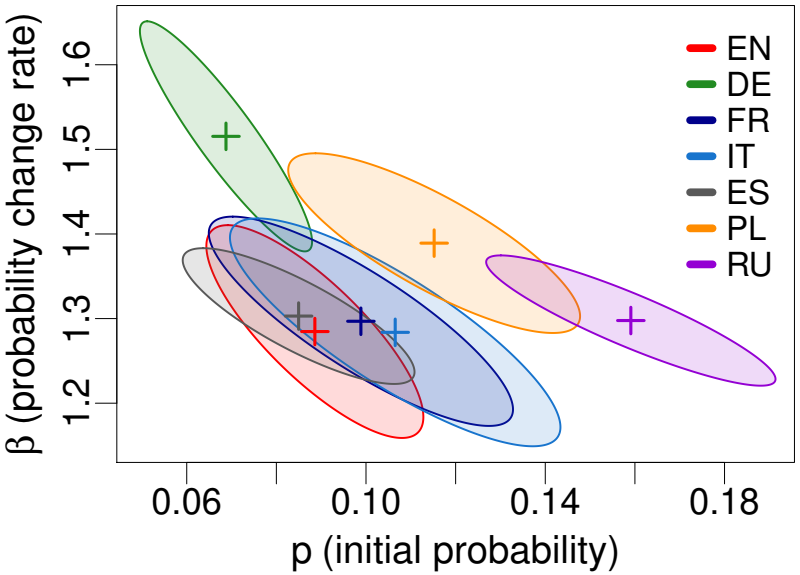
Typical values of the discrete Weibull distributions' parameters p and β are slightly different in different languages – each language has its own "region" on a p, β plane.



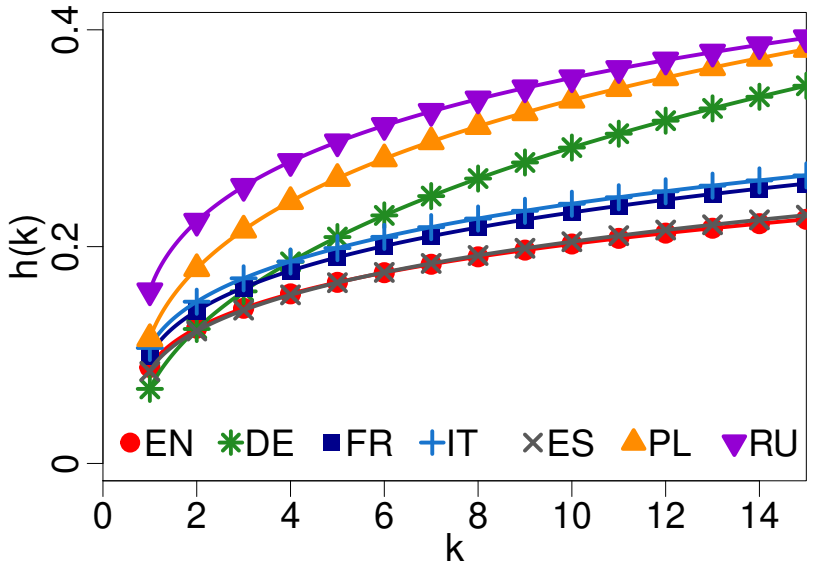
scatterplots of p and β for texts in different languages;
each plot (except the first) corresponds to one language, each point on a plot is one text

Differences between languages

Ellipses representing the regions in a (p, β) plane in which texts in different languages are concentrated



Hazard functions with p and β corresponding to ellipses' centers



- The average values of p and β (centers of the ellipses) can be used to determine "averaged" hazard function for each language.
- Within the range from $k \in [1; 15]$ (containing about 95% of the observed distances) the languages with the greatest values of h are Polish and Russian. This is partly due to the fact that the other studied languages utilize articles (words like *the*, *a*, *an* in English or *der*, *die*, *das* in German), which increase the distances and are absent in Polish and Russian.
- The lowest values of h can be attributed to English and Spanish. These are the two languages with the greatest numbers of speakers in the studied set of languages. Whether these two facts are somehow related - is an open question.

Summary

- The distribution of distances between consecutive punctuation marks in written texts can be approximated by discrete Weibull distribution.
- The parameters of the distribution, p and β, can be interpreted as parameters of a simple random process proposed as a mechanism behind the observed distribution of punctuation.
- Each language has its own typical patterns of punctuation usage, manifested by the average values of p and β.

References

- Stanisz, T., Drożdż, S., Kwapień, J., *Universal versus system-specific features of punctuation usage patterns in major Western languages*, Chaos, Solitons & Fractals (2023)

- Stanisz, T., Drożdż, S., Kwapień, J., *Complex systems approach to natural language*, to be published in Physics Reports