Zipf and Related Scaling Laws. 2. Literature Overview of Applications in Linguistics

Giedrė Būdienė [a] , Alytis Gruodis
Vilnius Business College, Kalvarijų str. 125, Vilnius, Lithuania

Received 1 February 2012, accepted 25 February 2012

Abstract. The overview of applications of Zipf and related scaling laws in linguistics are presented where the mathematical formulation of task in the framework of one and multi-dimensional distribution takes place. The object of quantitative linguistics represents the natural (western and eastern) as well as artificial languages (programming languages and random texts). Significant part of applications is devoted to the artificial intelligence systems based on zipfian distributions which are useful for cognitive search mechanisms.

Citations: Giedrė Būdienė, Alytis Gruodis. Zipf and Related Scaling Laws. 2. Literature Overview of Applications in Linguistics – *Innovative Infotechnologies for Science, Business and Education,* ISSN 2029-1035 – **1(12)** 2012 – Pp. 17-26.

Keywords: Zipf law; power law; quantitative linguistics; mathematical linguistics.
Short title: ZIPF law - linguistics - 2.

Introduction

Classification and ordering of selected sets constructed using multiple objects represent an unresolved problem in many areas of urban as well as scientific activity. Power law distributions represent statistical behaviour of classification in order to reselect the frequently used items from random occurred ones. Human speech belongs to one of irregular item distribution, and automatized language recognition (OCR, handwriting recognition, spelling correction etc) as well as artificial intelligence (augmentative communication, chat-boots etc) are based on statistical properties of language.

As a subdiscipline of *general linguistics*, the *quantitative linguistics* (or so-called *mathematical linguistics*) studies the quantitative aspects of structure typical for natural language. Static and dynamical approaches (present status and time-domain) allows understanding the changes in language morphology and undercrossing of several languages in certain field. It is obviously that statistical mathematical methods formulates the models, which are applicable by analysing the natural languages. Formulation of language laws allows to extrapolate the generalities of language into affinity group. Specific terms are used in such type modeling.

Linguistic object represents any text in any language where morphological, semantic and lexical rules were used in order to represent the certain idea.

Item represents an linguistic unit (smallest linguistic object). According to the most popular approach, item corresponds to single word (including all word forms). Another approaches allow to use two-words, three-words, also letters, syllables, morphological or semantic constructions etc. Taking more generally, any combination of *lexemes* (so called "base" words or dictionary-entries) extracted from regular or random texts according to certain rule could be treated as an item. For random text generation, the set of any letters of finite amount also could be treated as an item.

Token as the semantic element of programming or natural language could represent an linguistic unit and could be treated as an item.

Corpus represents structured set of texts (usually electronically stored, large amount) devoted for statistical analysis.

Our previous publication [1] was devoted to the overview the applications of Zipf and related scaling laws in economics. This work is aimed to the overview it in linguistics. We have selected about seventy typical references (up to 2011) including several of historic importance. Three approaches of the mentioned problem are presented below.

1. Mathematical formulation of task in the framework of one- and multi-dimensional distribution; description of the object and related laws in quantitative linguistics; description of models for word frequency distributions.
2. Zipfian applications for natural (western and eastern) as well as artificial languages (programming languages and random texts); principles of formation dictionaries.
3. Artificial intelligence systems based on ranked item frequency; cognitive mechanisms including search; language evolution as an informational process.

Quantitative linguistics is empirically based on the results of language statistics through statistics of any linguistic object.

[a]Corresponding author, email: *giedre@kolegija.lt*

1. Mathematical formulation of task

Zip law. George Zipf found the power-law-like word frequency dependence on word rank. The most frequent word appears twice as often as next most popular word, three times as often as 3rd most popular, and so on. So called zipfian distribution relates frequency $f(r)$ of item occurrence in finite corpus to item rank $r(w)$ according to Eq. (1).

$$f(r) = \frac{\alpha}{r^{\gamma}} \tag{1}$$

$$\log f(r) = \alpha - \gamma \cdot \log r \tag{2}$$

In particular case for Zipf distribution, exponent $\gamma \approx 1$. In a logarithmic scale, this dependence represents a straight line, graphical charts are presenten in previous Ref. [1].

For finite corpus of size N, the Zipf coefficient K_{lan} depending on language could be established according to Eq.(3):

$$K_{lan} = N\frac{r(w)}{c(w)} \tag{3}$$

where $c(w)$ - number of selected ranked items w. The simplest case of Zipf law is the famous f^{-1} hyperbolic function. For detailed textbook, see very large study [2] prepared by Saichev et al.

Li in review article [3] devoted to the 100th anniversary of the birth of George Zipf accented the ubiquity of Zipf law. All questions are not answered yet.

1. Is there a rigorous test in fitting real data to Zipf law?
2. In how many forms does Zipf law appear?
3. In which fields are the data sets claiming to exhibit Zipf law?

Heaps law and Herdan law. Harold Heaps discovered distribution of vocabulary size on text length [4]. Number of distinct items (words) $V(n)$ in a part of text containing n items is exponentially proportional to n - so called **Heaps** law, see Eq.(4).

$$V(n) = \alpha \cdot n^{\gamma} \tag{4}$$

With English text corpora, typically $\alpha \in [10 \div 100]$, and $\gamma \in [0.4 \div 0.6]$. Heaps law is asymptotically equivalent to Zipf law concerning the frequencies of individual items (words) within a text.

Gustav Herdan [5] proposed following formulation: the logarithm of vocabulary size divided by the logarithm of text size is a constant smaller than 1 – so called **Herdan** law, see Eq.(5). It is evident that Eq.(5) corresponds to Eq.(4) when α=1.

$$\gamma = \frac{\log V(n)}{\log n} \tag{5}$$

Task of item classification could be treated as a significant part in signalling theory, which examines communication types between individuals. The nature of the Zipf and Heaps laws is not yet clear. According to statements in signalling theory, the language items distributions via power law expressions seem to be specific for natural as well as artificial languages only, otherwise, another signal systems, based on-demand assumption, show stochastic behaviour only [6].

Alfred Lotka states that the number of authors making n contributions is proportional to the n^{-2}. **Lotka law** describes the frequency of publication $V(n)$ by authors n in any given field. It could be treated as one of a variety of Zipf law (γ=2).

$$V(n) = \frac{\alpha}{n^2} \tag{6}$$

Egghe [7] investigates Herdan law and Heaps law from a purely mathematical and informetric point of view. Dependencies according to Lotka law (exponent γ=2) and Zipf law (exponent γ=1) must be treated as expression of boundary conditions by analysing text in linguistics (citations and regular text, respectively).

Bernhardsson et al. [8] analyse text-length dependence of the power-law index of a single book. They have been found that exponent value decreases from 2 to 1 with increasing text length according to extended Heaps law and Zipf law respectively. Authors proposed an idea that the systematic text-length dependence can be described by a meta book concept, which is an abstract representation reflecting the word-frequency structure of a text.

Analysing on any text usually starts from two operations.

1. Calculating of item frequency distribution on rank. In many cases, Zipf or Lotka dependences are expected, for example, see Fig. 1. Exponent $\gamma \in [1 \div 2]$.
2. Calculating of vocabulary size distribution on text size. As usually, Heaps dependence expected, for example, see Fig. 2.

Fig. 3 represents the plot (in log-log coordinates) of ranked word frequency. English corpus was obtained from Wikipedia (data until November 27, 2006) [9]. As expected for English language, the most popular words are "the"(r=1), "of"(r=2), "and"(r=3), also "a", "in", "be", "to", "that" and so on. Actually, Zipf law represents an harmonic series r^{-1}, which describes the real word distribution quite well in first assumption only.

Initial part of dependence in interval AB is claimed as non-zipfian dependence ($\gamma \approx 0.5$). Part AB represents dependence of the words which are morphologically or semantically requested (according to the language construction).

Middle part of dependence in interval BC represents Zipf law (γ=1).

Part CD represents dependencies of rarely used words, so called citation words according to Lotka law (exponent γ=2). Also long tail dependence in interval CE represents Zipf-Mandelbrot law ($\gamma \geq 2$).

These lines correspond to three distinct parameterizations of the Zipf-Mandelbrot distribution.

Models for Word Frequency Distributions. It has taken more than 100 years to discuss the item frequency occurrence in different sciences and linguistic areas such as natural sciences (particle distribution, gene sequence, and earthquakes), economics (market parameters, growth prognosis),

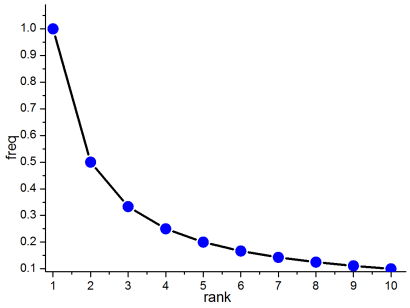

Fig. 1. Zipf law: simulated word frequency distribution on word rank according to Eq.(1), α=1, γ=1.

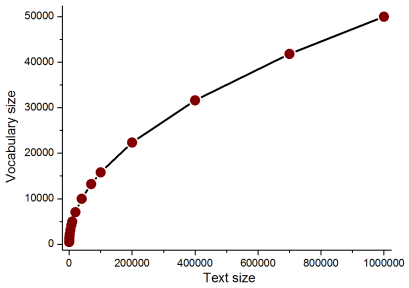

Fig. 2. Heaps law: simulated vocabulary size distribution on text size according to Eq.(4), α=50, γ=0.5.

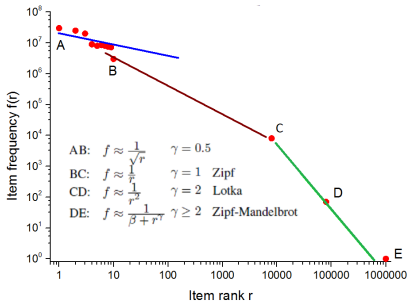

Fig. 3. Ranked word frequency dependence (in log-log scale). English corpus from Wikipedia, Nov 27, 2006. Adapted according to Ref. [9].

urbanistics (the dynamical sizes of cities) etc. Linguistics represents a specific field area in information exchange where chaotic and semi-chaotic sequences – items are encompassed into general power laws. Zipf law uncovers the relationship between word frequency $f(r)$ and its rank r – see Eq. (1).

Zipf [10] states that the *Principle of Least Effort* as the primary principle governs our entire individual and collective behaviour of all sorts, including the behaviour of our language and preconceptions. Kirby [11] analyses validity of Zipfs law using different type of examples. Simulation of organisational behaviour gives the zipfian dependence.

Kosmidis et al. [12] probed to formalize the language system using a simple expression for the Hamiltonian, which is directly implied by the Zipf law. Several language properties such as universality of the Zipf exponent, the vocabulary size of children, the reduced communication abilities of people suffering from schizophrenia could be able to explain.

Historically most important Zipf law according to Eq.(1) and several derived/related laws such as **Paretto distribution** according to Eq.(7) can be applied for strong selection, sorting, prediction, recognition of linguistic items of different languages.

$$f(r) = \left[\frac{r}{r_{min}}\right]^{-\gamma} \tag{7}$$

Egghe [13] analyses graphically the relation between the fraction of the items and the fraction of the sources producing these items. Paretto distribution or so called 80/20-rule by fitting Lorenz curve is evident. Egghe claims that the share of items as a function of the corresponding share of sources increases with increasing size of the system.

Newman [14] reviews some of the empirical evidence for the existence of power-law forms and the theories proposed to explain them. The origin of power-law behaviour has been a topic of debate for more than a century. Darooneh [15] analyses the statistics of ranked words in natural languages using the rank-frequency plot of these words. In that case, model of fractional brownian motion was used in order to improve the power law prediction. Verification of a finite size scaling ansatz was done. This routine allows finding the correct relation between the Zipf exponent and the Hurst exponent characterizing the fractional brownian motion.

Montemurro [16] proposes the revisited **Zipf–Mandelbrot law** in the context of linguistics - see Fig. 4. Its well known that Zipf–Mandelbrot law describes the statistical dependence of the items from certain corpus only. Significant deviations become statistically relevant as larger corpora are considered.

$$f(r) = \frac{\alpha}{(1+\beta r)^{\gamma}} \tag{8}$$

$$f(r) = \frac{\alpha}{\beta + r^{\gamma}} \tag{9}$$

By varying the scale parameter β, it is possible to fit the word frequency dependence in part AB - see Fig.3.

Lognormal law. Fig. 5 represents the probability density function of a log-normal distribution according to Eq.(10). Mentioned non-symmetrical multi-dimensional distribution contains location parameter μ and scale parameter σ.

$$f_X(x,\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] \tag{10}$$

Weibul distribution. Fig. 6 represents the probability density function of a Weibull random variable which is useful function in order to simulate particle size distribution. Eq.(11) represent expression for ($r \leq 0$), where positive shape and scale parameters take place: k>0, λ>0.

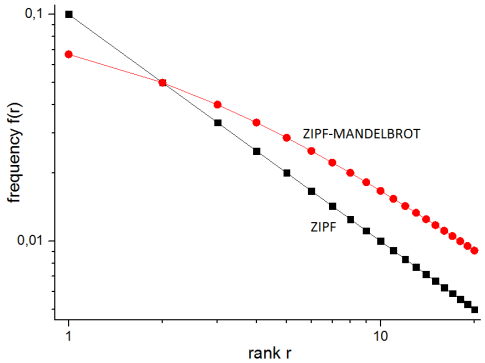

Fig. 4. Simulated word frequency dependence on rank.
Zipf distribution according to Eq.(1), black;
Zipf-Mandelbrot distribution according to Eq.(8), β=0.5, red.
For both curves, α=0.1, γ=1.

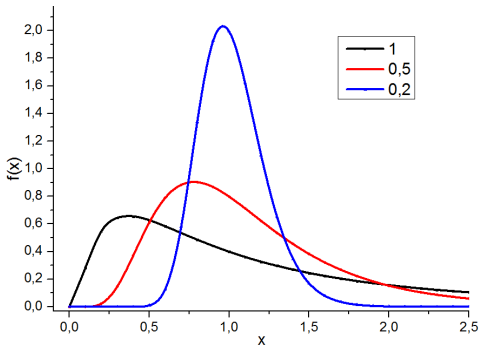

Fig. 5. Log-normal distribution according to Eq.(10), when μ=0. Scale parameter σ=1, black; σ=0.5, red; σ=0.2, blue.

$$f(r,\lambda,k) = \frac{k}{\lambda}\left(\frac{r}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{r}{\lambda}\right)^{k}\right] \qquad (11)$$

Baayen [17] describes three models for word frequency distributions: generalized inverse Gauss-Poisson law; log-normal law according to Eq.(10) and Zipf law according to Eq.(1). Goodness of fit and rationale are commented. It was concluded that no model could vouch the exclusive validity. The role of morphology in shaping word frequency distributions is actual because the vocabulary richness in literary studies correlates to the morphological productivity in linguistics.

Traditionally, zipfian dependencies are devoted for ranked distributions in linguistics, bet otherwise, full statistical analysis such as ANOVA allows analysing the distributions of random and quasi-random items. Limpert et al. [18] analyse applicability of log-normal distribution in several areas of science.

Mitzenmacher [19-20] represent quite brief history of applications related to the several recently proposed models such as lognormal and power law distributions in several areas including validation of models and control of systems. Analysis in many field allow concluding that lognormal distributions have arisen as a possible alternative to power law distributions across many fields.

Menzerath-Altmann law is devoted for certain linguistic construction which contains the constituents: the size of the constituents decrease with increasing size of the construction. Eq.(12) relates $f(r)$ syllable length to r - number of syllables per word. Kohler [21] suggested that linguistic segments contain information about its structure (besides the information that needs to be communicated).

$$f(r) = \frac{\alpha \cdot r^{\beta}}{\exp(\gamma r)} \qquad (12)$$

Aren law (exponential distribution according to Eq.(13)) plays significant role in process modeling. Aren expression could be derived as the special case of Menzerath-Altmann law, when β=0.

$$f(r) = \frac{\alpha}{\exp(\gamma r)} \qquad (13)$$

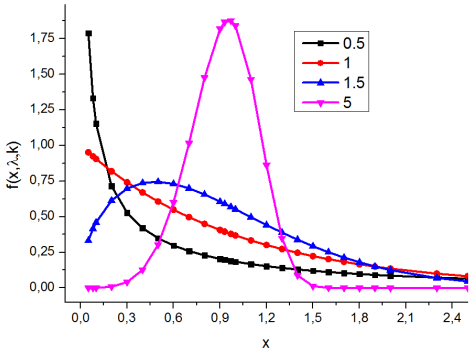

Fig. 6. Weibul distribution according to Eq.(11) with different exponent k={0.5, 1, 1.5, 5} when λ=1.0.

Eliazar et al. [22] presented a universal mechanism for the temporal generation of power-law distributions with arbitrary integer-valued exponents. Hill [23] describes several approaches for applicability of power-law. Generally, deviations from the one-exponential distribution cover the stochastic manifestation of item groups, and for such case data fitting must be done using sophisticated models: **Yule-Simon** distribution according to Eq.(14) or **beta function**, so called Euler integral according to Eq.(15).

$$f(r) = \frac{\alpha \cdot \beta^{r}}{r^{\gamma}} \qquad (14)$$

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t. \qquad (15)$$

Li et al. [24] describe several two-parameter models, including beta function, Yule function, Weibull function for linguistic analysis. Letter frequencies, word-spacing, and word frequencies were used as the ranked linguistic data. Li claims that beta function fits the ranked *letter* frequency distribution the best, but otherwise, Yule function fits the ranked *word-spacing* distribution the best. Altmann, beta, Yule functions all slightly outperform the Zipf power-law function in word ranked- frequency distribution.

Naumis et al. [25] probed to solve the task of fit rank distributions when fits usually fail at the tail. Naumis proposed to use beta-like function. Authors claim that the observed behaviour at the tail seems to be related with the onset of different mechanisms that are dominant at different scales, providing crossovers and finite size effects.

2. Zipfian applications for languages

Dictionaries. Standardization of language always starts from dictionaries made up from corpuses. Formulated from first suggestions as an empirical law, Zipf law represents a universal distribution for natural language not depending on chosen language, word quantity in item, and specific sphere of language usage. Universality of law allows to use it in digital linguistics. Maslov [26] analyses the specific usage of power laws: how linguistic applications could be applied for

dictionary forming. There are several ground principles necessary for resource effectively compiling in order to create the time-saving search. On one hand, dictionaries as specific word sorting form represent a top-application of Zipf law in linguistics; on the other hand, time-domain processes of word item occurrence in sequences cannot be recognized from different slopes of Zipf curves. Willys [27] states that king of ambiguity does not allow us to solve the reverse task. Maslov et al. [28] analysed the big number of formulas of linguistic statistics. The notions of real and virtual cardinality of a sign were introduced. It was concluded that formula refining Zipf law for the occurrence frequencies in frequency dictionaries can be extended to the big semiotic systems. Wyllys [29] provided the jargon stardartization routine in scientific writing using zipfian distributions.

Zanette [30] analyses the family name distribution as time-dependent process. The evolution of family-name distributions is limited vertically depending on cultural features. Empirical power law distribution takes place.

Murtra et al. [31] analyse the applications of Zipf law in the context of a very general class of stochastic systems. Complexity of the description of the system provided by the sequence of observations is the one expected for a system evolving to a stable state between order and disorder. Powers [32] demonstrate how Zipf analysis can be extended to include some of the phenomena not explainable using power-law distributions.

Natural western languages. Regularities of zipfian item distribution are confirmed in many modern languages of great importance, for example American English [33]. Maslov [34] analyses English, French, Spanish languages as the most dominating western languages by expansion among spoken population all over the world. Typical application of Zipf distribution is related to the forming of frequency dictionary. It should be considered that varieties of dictionary might be defined as a logarithmic correction to the Zipf–Mandelbrot law whereas the main problem lies in the tails of distribution. The tails are formed from less-frequently or seldom occurring words derived or constructed morphologically according to the informal rules of spoken and written language.

Tuzzi et al. [35] analyse nonstandard Italian texts such as official presidential speeches in order to confirm the Zipf law. The results showed the unique lexis of the corpus. The analyses allow us to find a position for each president on the synthetism/analytism scale and individual characteristic features of each president.

Popescu et al. [36] presented novel method for language analysing. Even though Zipf law can be applied to a variety linguistic data, a common formula of law cannot be derived to be applicable to the all data sets. New approach to the problem consisting of the multi-component analysis was proposed and tested in 20 languages.

Ha et al. [37] analyse extremely large corpuses: English corpus of 500 million word tokens and 689,000 word types. Is was established the zipfian dependency takes place: the usual slope close to γ=1 for rank less than 5,000, but then for a higher rank it turns to give a slope close to γ=2. Ha concludes that presented phenomenon is done due to foreign words and place names. The Zipf curves for Celtic, Irish languages were presented. Because of the larger number of word types per lemma, it remains flatter than the English curve maintaining a slope of γ=1 until a turning point of about rank 30000.

Ausloos [38] described translation problem: a comparison of two English texts also translated into Esperanto are discussed in order to observe whether natural and artificial languages significantly differ from each other. Word frequencies distribution (studied by a Zipf method) and word lengths distribution (studied by a Grassberger–Procaccia technique) were used. Quantitative statistical differences between the original English text and its Esperanto translation were found. Different power law distributions were observed. The Zipf exponent is equal to $\gamma \in [0.50 \div 0.30]$ depending on how a sentence is defined. Together with the attractor and space dimension, such parameters could also be attached for measurement of the author style versatility.

Programming languages. Zhang [39] discovered the power-law regularities in the distribution of lexical tokens in modern Java, C++ and C programs. It was established that such distributions follow Zipf–Mandelbrot law, and the growth of program vocabulary follows Heaps law.

Natural eastern languages. Natural eastern languages are typical examples of expansive processes of language formation in comparison with western languages. Dahui et al. [40] presented the research where data of traditional and modern Chinese literature was used. Significant differences between Zipf law distributions of mentioned Chinese character sets were found - due to disordered growth of dictionary. Dahui established that the true reason for Zipf law in language is that growth and preferential selection mechanism of word or character in given language.

Ranking problems occur when parallel texts in Chinese and English are analysed according to the frequency distribution. Zipf distribution is applicable until certain barrier of token amount (1 thousand for Chinese and 5 thousand for English). Presence of barrier can be explained by excess of additional tokens, which were put into the context as semantically uncompleted forms. Ha et al. [41] state that when single are combined together with n-gram characters in one list and put in order of frequency, the frequency of tokens in the combined list follows Zipf law - $\gamma \approx 1$. This unexplained behaviour is also found for English 2-byte and 3-byte word fragments.

Xiao [42] analyses applicability of Zipf Law in Chinese word frequency distribution. It was also found out that low frequency words constitute over half of the corpus word occurrences. This is the main reason why data sparse in statistical approaches could not be significantly reduced even ex-

panding corpus scale.

Sen et al. [43] solved the task of validity of Zipf law related to the word (item) length and the frequency was confirmed by analysing the big sets (up to 5,800 words). The main exception is found to be one-letter words.

Changing object of investigation from regular token to specific items – family names – it is necessary to describe the complicated origin of item, which encompasses family name as well as birthplace. Family name distributions with or without the information of the regional origins are applicable to power function - Zipf law. Kim et al. [44] and Miyazima et al. [45] presented the analysis family names belonging to Korean and Japan societies, respectively. In addition, Miyazima states that the relation between size and rank of a family name also shows a power law. Yamada et al. [46] used another fitting technique by means of q-exponential function for the distribution of Japanese family names in order to obey power-law distribution (Zipf law).

Several differences between phonogram-based language (English) and ideogram-based language (Japanese) were found by analysing power law distribution by Nabeshima et al. [47]. It was established that frequency of word usage against rank follows power-law function with exponent γ=1 and, for Japanese ideogram, it follows stretched exponential (Weibull distribution) function.

Sheng et al. [48] analyse the statistical properties of English and Chinese written human language. New approach instead of power law distribution was used: so called *framework of weighted complex networks*. These observations indicate that the two languages may have different linguistic mechanisms and different combinatorial natures. The results display some differences in the structural organizations between the two language networks.

Natural language imitation through random text. Randomly generated texts (RGT) represent sets of items with different probability. Distributions of item frequencies of RGT and English are similar and complies with Zipfs law. Li [49] claims that frequency of occupancy of a word is almost an inverse power law function of its rank and the exponent of this inverse power law is very close to γ=1.

Several methods of text generating could be presented such as *intermittent silence process*. Cancho [50] argued that the real power-law type distribution of word frequencies could be explained by generating a random sequence of characters by means of *intermittent silence process*. According to such method, expected frequency spectrum and the expected vocabulary size as a function of the text size could be efficiently calculated.

Monkey-at-the-typewriter model. Perline [51] describes the application of the classical Mandelbrot *monkey-at-the-typewriter* model as the model where Zipf inverse power law is applicable. An explicit asymptotic formula for the slope of the log-linear rank-size law in the upper tail of this distribution is also obtained. By usage of the same *monkey-at-the-typewriter* model, Conrad et al. [52] showed so called recent confusion, where the rank-frequency distribution follows a lognormal distribution. This special model arises in particular case, where letters are hit with unequal probability.

On the other hand, Cancho [53] demonstrate by means of three different statistical tests that ranks derived from random texts and ranks derived from real texts are statistically inconsistent. Cancho concludes that the good fit of random texts to real Zipf law-like rank distributions has not yet been established.

3. Artificial intelligence systems

Cognitive mechanisms including search. Serrano et al. [54] studied the written text problem in the context of text recognition tasks. Two approaches were used for modeling: Zipfs law and Heaps law. It was established the significant relation between the burst nature of rare words and the topical organization of texts. The dynamic word ranking and memory across documents – such two factors could be treated as a key mechanisms explaining the non trivial organization of written text.

Wyllys [55] analyses implications of Zipf law for the design of information systems. He claims that only vocabulary control could be done using Zipf law. Wyllys says that sentence about universality of Zipf law (that different subject-fields may be characterized by different slopes of Zipf curves) seems to have no practical applications in information system design at present (may be in future).

Blanchard [56] solves the problem of a document retrievals in patent mapping tools. Previous stopword list technique was used – as a system which modified the retrieval words into more powerful (i.e. they dramatically impacts the final output and analysis). Stopword lists depend on the document corpus analysed according to power-law.

Calderon et al. [57] analyse the distribution of words in Spanish texts of Latin-American writers from Zipf law perspective. New approach to Zipf law dependencies was used: the frequency of repetition of a particular word among other different words was analysed in order to solve the linguistic problem using statistical approach.

Kello et al. [58] analyse linguistic activities using scaling laws which suggest the existence of patterns that are repeated across scales of analysis. Variable can vary in region between several types. In that case recurrence of scaling laws has prompted a search for unifying principles. In language systems, scaling laws can reflect adaptive processes of various types and are often linked to complex systems near critical points. Findings of scaling laws in cognitive science are indicative of scaling invariance in cognitive mechanisms.

Caron et al. [59] analyse semantic extraction of word groups belonging to the different regions of interest. Zipf law and inverse Zipf law were used in order to characterize the structural complexity of image textures. The distribution

of pattern frequency was modeled as power law distributions. Method allows the detection of regions of interest, which are consistent with human perception, where inverse Zipf law is particularly significant.

Altmann et al. [60] analyse big corpuses where the language has different levels of formality. These distributions are well characterized by a stretched exponential (Weibull) scaling. Distributions of distances between successive occurrences of the same word display some deviations from a Poisson process. The extent of this deviation depends strongly on semantic type. A generative model of this behaviour that fully determines the dynamics of word usage was developed.

Automatic text analysis is grounded on Luhn assumption [61] that frequency data can be used to extract words and sentences in order to represent a certain document. Losee [62] analyses regularities in the statistical information provided by natural language terms about neighbouring terms. We find that when phrase rank increases, moving from common to less common phrases, the value of the expected mutual information measure (EMIM) between the terms regularly decreases. Luhn model suggests that mid-range terms are the best index terms and relevance discriminators. Interpretation of Zipf law from information theoretic point view was provided. Using the regularity noted above, we suggest that Zipf law is a consequence of the statistical dependencies that exist between terms, described here using information theoretic concepts.

New teaching/learning methods. Vousden [63] uses application of Zipf law in order to choose the English teaching material as spelling-to-sound units. In that case, the quantity and adaptability could be rationalized in high degree. Alexander et al. [64] use application of Zipf law for helping the students to create the interconnection between mathematics and other disciplines.

Language evolution as an informational process. In quantitative linguistics, **Piotrowski** law [65] describes the process of language change through several parameters:

i) vocabulary growth;
ii) the dispersion of foreign or loan words;
iii) changes in the inflectional system etc.

Initial hypothesis (everything in language changes as a result of interaction between old forms and new forms) could be formulized through differential equation:

$$\frac{\mathrm{d}p_t}{\mathrm{d}t} = k_t \cdot p_t \cdot (C - p_t) \tag{16}$$

where $\mathrm{d}p_t$ - change in the proportion; p_t - proportion of new forms; k_t - time-dependent function.

Most important solution of mentioned differential equation is presented below. In case, if C=1 and k_t=b, solution represents so-called **logistic** curve for modeling the growth phenomena (α is the integration constant). Fig. 7 represents the

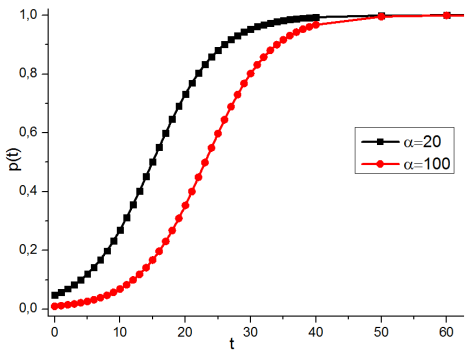

Fig. 7. Logistic distributions for growth modeling according to Eq.(17). β=0.2.
Different integration constant $\alpha \in \{20; 100\}$.

logistic curve made up by Eq.(17).

$$p(t) = \frac{1}{1 + \alpha \exp(-\beta t)} \tag{17}$$

Joshua et al. [66] use mathematical models to analyse the major transitions in language evolution. Word-formation is described as a process related to Shannon noisy coding theorem. Model of the population dynamics of words and the adaptive emergence of syntax is present.

Bernhardsson et al. [67] analyse functional form of the word-frequency distribution. So called null model was used where the words are randomly distributed throughout the text. Initial assumption of sharing characteristics (real novel shares many characteristic features with a null model) was used together with second (functional form of the word-frequency distribution of a novel depends on the length of the text in the same way as the null model). This means that an approximate power-law tail will have an exponent which changes with the size of the text-section which is analysed. The size-transformation of a novel is found to be well described by a specific Random Book Transformation.

Shenon entropy in evolution model. Maillart et al. [68] studied the evolution processes of open source software projects in Linux distributions, which offer a remarkable example of a growing complex of self-organizing adaptive system. The ingredients of stochastic growth models were established empirically which are previously conjectured to be at the origin of Zipf law.

Unpredictability of information content could be characterized by Shannon entropy H where $P(x)$ is the probability that variable X occupies the state x. Summation must be provided over all states N.

$$H(X) = -\sum_{i=1}^{N} P(x_i) \cdot \log_2(P(x_i)) \tag{18}$$

Dover [69] proposed novel formalism of maximum principle of Shannon entropy in order to derive the general power law distribution function. There are big number of examples where Boltzmann entropy is related to the paradigm of

"internal order": complexion, self-interacting, self-organized system etc. Evolution of structure could be modeled by describing the noninteracting conditions since the Shannon entropy is equivalent to the Boltzmann entropy under equilibrium. This formalism was demonstrated in toy model where Zipf law comes out as a natural special point of the model.

Nesterova [70] presented large review of applications of Shenon entropy. Main paradigms - system, structure, information - and corresponding parameters - entropy, negentropy - are desribed for characterization two different - metric as well as information system.

Cancho [71] describes a general communication model where objects map to signals, a power function for the distribution of signal frequencies is derived. Cancho claims that many systems in nature use non-trivial strategies for easing the interpretation of a signal. Presented model relies on the satisfaction of the receiver communicative needs when the entropy of the number of objects per signal is maximized. Estimation in linguistic context is surprising: present exponent ($\gamma\approx2$) is clearly different from the typical of Zipf law ($\gamma\approx1$). It means that Zipf law reflects some sort of optimization. On other hand, the words are used according to the objects (e.g. meanings) they are linked to (linguistic approach).

Cancho [72] analyses the new model for Zipf law proposed for the human word distribution in the framework of information theory: from a no communication phase to a perfect communication phase. Scaling consistent with Zipf law is found in the boundary between phases. The exponents are consistent with minimizing the entropy of words. Presented model is especially suitable for the speech of schizophrenics. Zipf exponent predicted for the frequency versus rank distribution is in a range where γ>1, which may explain the word frequency distribution of some schizophrenics and some children, with $\gamma\in[1.5\div1.6]$. Among the many models for Zipf law, none explains Zipf law for that particular range of exponents. In particular, two simplistic models fail to explain that particular range of exponents: intermittent silence and Simon model.

Conclusion

1. Many linguistic ranked item frequency distributions could be described using Zipf or Zipf-Mandelbrot law with exponent $\gamma\approx1$. Increasing of exponent up to $\gamma\approx2$ (long tail problem) is related to the stochastic nature of items.
2. Yule, beta and Manzerath-Altman distributions could be treated as the "modifications" of more general power-low where specific fitting parameters are useful for precisely adequacy to original distribution.
3. In linguistics, power-law represents influence of human behaviour where language as a communication tool can be used. Dependencies according to Lotka law (exponent γ=2) and Zipf law (exponent γ=1) must be treated as expression of the boundary conditions by analysing text in linguistics.

References

1. Artūras Einikis, Giedrė Būdienė, Alytis Gruodis. Zipf and Related Scaling Laws. 1. Literature Overview of Applications in Economics. – *Innovative Infotechnologies for Science, Business and Education* ISSN 2029-1035 – 2(11) (2011) 27-36.
2. Alexander I. Saichev, Yannick Malevergne, Didier Sornette. Theory of Zipf's Law and Beyond. – Berlin: Springer, 2010.
3. Wentian Li. Zipf's Law Everywhere. – *Glottometrics* 5 (2002) 14-21.
4. Heaps Harold Stanley (1978), Information Retrieval: Computational and Theoretical Aspects, Academic Press. – Heaps law is proposed in Section 7.5 (pp. 206–208).
5. Leo Egghe. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments. – *Journal of the American Society for Information Science and Technology* 58(5) (2007) 702–709.
6. Rebecca Bliege Bird, Eric Alden Smith. Signaling Theory, Strategic Interaction, and Symbolic Capital. – *Current Anthropology* 46(2) (2005) 221-248.
7. Leo Egghe. Untangling Herdan's Law and Heaps' Law: Mathematical and Informetric Arguments. – *Journal of the American society for information science and technology* 58(5) (2007) 702–709.
8. Sebastian Bernhardsson, Luis Enrique Correa da Rocha, Petter Minnhagen. The meta book and size-dependent properties of written language. – *New Journal of Physics* 11 (2009) 123015.
9. <http://en.wikipedia.org/wiki/Zipf%27s_law>, accessed 2012.01.15.
10. George Kingsley Zipf. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology – Addison-Wesley Press Inc., 1949.
11. Geoff Kirby. Zipf's law. – *UK Journal of Naval Science* 10(3) (1985) 180-185.
12. Kosmas Kosmidis, Alkiviadis Kalampokis, Panos Argyrakis. Statistical mechanical approach to human language. – *Physica A* 366 (2006) 495–502.
13. L. Egghe. The dependence of the height of a Lorenz curve of a Zipf function on the size of the system. – *Mathematical and Computer Modelling* 43 (2006) 870–879.
14. M.E.J. Newman. Power laws, Pareto distributions and Zipf's law. – *Contemporary Physics* 46(5) (2005) 323-351.
15. A.H. Darooneh, B. Rahmani. Finite size correction for fixed word length Zipf analysis. – *Eur. Phys. J. B* 70 (2009) 287–291 .
16. Marcelo A. Montemurro. Beyond the Zipf–Mandelbrot law in quantitative linguistics. – *Physica A* 300 (2001) 567–578.

17. Harald Baayen. A Linguistic Evaluation. – *Computers and the Humanities* 26 (1993) 347-363.
18. Eckhard Limpert, Werner A. Stahel, Markus Abbt. Log-normal Distributions across the Sciences: Keys and Clues. – *BioScience* 51(5) (2001) 341-352.
19. Michael Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. – *Internet Mathematics* 1(2) (2003) 226-251.
20. Michael Mitzenmacher. The Future of Power Law Research. – *Internet Mathematics* 2(4) (2005) 525-534.
21. Reinhard Köhler. Zur Interpretation des Menzerathschen Gesetzes. – *Glottometrika* 6 (1984) 177–183.
22. Iddo Eliazar, Joseph Klafter. Temporal generation of power-law distributions: A universal 'oligarchy mechanism'. – *Physica A* 377 (2007) 53–57.
23. Bruce M. Hill. The Rank-Frequency form of Zipfs law. – *Journal of the American Statistical Association* 69(384) (1974) 1017-1026.
24. Wentian Li, Pedro Miramontes and Germinal Cocho. Fitting Ranked Linguistic Data with Two-Parameter Functions. – *Entropy* 12 (2010) 1743-1764.
25. G.G. Naumis, G. Cocho. Tail universalities in rank distributions as an algebraic problem: The beta-like function. – *Physica A* 387 (2008) 84–96.
26. V. P. Maslov. Quantum Linguistic Statistics. – *Russian Journal of Mathematical Physics* 13(3) (2006) 315–325.
27. Ronald E. Wyllys. Empirical and Theoretical Bases of Zipf's Law. – *Library Trends* 30(1) (1981) 53-64.
28. V. P. Maslov, T. V. Maslova. On Zipf's Law and Rank Distributions in Linguistics and Semiotics. – *Mathematical Notes* 80(5) (2006) 679–691. – Translated from *Matematicheskie Zametki* 80(5) (2006) 718–732.
29. Ronal Eugene Wyllys. The measurement of jargon standardization in scientific writing using rank-frequency ("Zipf") curves. PhD thesis. – University of Wisconsin, 1974.
30. Damisan H. Zanette, Susanna C. Manrubia. Vertical transmission of culture and the distribution of family names. – *Physica A* 295 (2001) 1–8.
31. Bernat Corominas-Murtra, Ricard V. Sole. Universality of Zipf's law. – *Physical Review E* 82 (2010) 011102.
32. David M. W. Powers. Applications and Explanations of Zipfs Law. – In: D. M. W. Powers (ed.) NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning – ACL, 1998. – Pp. 151-160.
33. H. Kučera, W. N. Francis. Computational Analysis of Present-Day American English. – Brown University, 1967.
34. V. P. Maslov. The Lack-of-Preference Law and the Corresponding Distributions in Frequency Probability Theory. – *Mathematical Notes* 80(2) (2006) 214–223. – Translated from *Matematicheskie Zametki* 80(2) (2006) 220–230.
35. Arjuna Tuzzi, Ioan-Iovitz Popescu, Gabriel Altmann. Zipf's Laws in Italian Texts. – *Journal of Quantitative Linguistics* 16(4) (2009) 354–367.
36. Ioan-Iovitz Popescu, Gabriel Altmann, Reinhard Köhler. Zipf's law—another view. – *Qual. Quant.* 44 (2010) 713-731.
37. Le Quan Ha, Francis J Smith. Zipf and Type-Token rules for the English and Irish languages. – MIDL, Paris, 29-30 novembre 2004. – Pp. 65-70.
38. M. Ausloos. Equilibrium and dynamic methods when comparing an English text and its Esperanto translation. – *Physica A* 387 (2008) 6411–6420.
39. Hongyu Zhang. Discovering power laws in computer programs. – *Information Processing and Management* 45 (2009) 477–483.
40. Wang Dahui, Li Menghui, Di Zengru. True reason for Zipf's law in language. – *Physica A* 358 (2005) 545–550.
41. Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming and F. J. Smith. Extension of Zipf's Law to Word and Character N-grams for English and Chinese. – *Computational Linguistics and Chinese Language Processing* 8(1) (2003) 77-102.
42. Hang Xiao. On the Applicability of Zipf's Law in Chinese Word Frequency Distribution. – *Journal of Chinese Language and Computing* 18(1) (2008) 33-46.
43. B.K. Sen, Khong Wye Keen, Lee Soo Hoon, Lim Bee Ling, Mohd Rafae Abdullah, Ting Chang Nguan, Wee Siu Hiang. Zipf's law and writings on LIS. – *Malaysian Journal of Library & Information Science* 3(2) (1998) 93-98.
44. Beom Jun Kim, Sung Min Park. Distribution of Korean family names. – *Physica A* 347 (2005) 683–694.
45. Sasuke Miyazima, Youngki Lee, Tomomasa Nagamine, Hiroaki Miyajima. Power-law distribution of family names in Japanese societies. – *Physica A* 278 (2000) 282-288.
46. Hiroaki S. Yamada, Kazumoto Iguchi. q-exponential fitting for distributions of family names. – *Physica A* 387 (2008) 1628–1636.
47. Terutaka Nabeshima, Yukio-Pegio Gunji. Zipf's law in phonograms and Weibull distribution in ideograms: comparison of English with Japanese. – *BioSystems* 73 (2004) 131–139.
48. Long Sheng, Chunguang Li. English and Chinese languages as weighted complex networks. – *Physica A* 388 (2009) 2561-2570.
49. Wentian Li. Random Texts Exhibit Zipfs-Law-Like Word - Frequency Distribution. – *IEEE Transactions on information theory* 38(6) (1992) 1842-1845.
50. Ramon Ferrer-i-Cancho, Ricard Gavalda. The Frequency Spectrum of Finite Samples from the Intermittent Silence Process. – *Journal of the American Society for Information Science and Technology* 60(4) (2009) 837–843.
51. Richard Perline. Zipf's law, the central limit theorem, and the random division of the unit interval. – *Physical Review E* 54(1) (1996) 220-223.
52. Brian Conrad and Michael Mitzenmacher. Power Laws for Monkeys Typing Randomly: The Case of Unequal Probabilities. – *IEEE Transactions on information theory* 50(7) (2004) 1403.

53. Ramon Ferrer-i-Cancho, Brita Elvevag. Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution. – *PLoS ONE* 5(3) (2010) e9411. – <www.plosone.org>, accessed 2011.11.19.
54. M. Angeles Serrano, Alessandro Flammini, Filippo Menczer. Modeling Statistical Properties of Written Text. – *PLoS ONE* 4(4) (2009) e5372. – <www.plosone.org>, accessed 2011.11.19.
55. Ronald E. Wyllys. Empirical and Theoretical Bases of Zipf's Law. – *Library Trends* 30(1) (1981) 53-64.
56. Antoine Blanchard. Understanding and customizing stopword lists for enhanced patent mapping. – *World Patent Information* 29 (2007) 308–316.
57. F. Calderon, S. Curilef and M. L. Ladron de Guevara. Probability distribution in a quantitative linguistic problem. – *Brazilian Journal of Physics* 39(2A) (2009) 500-502.
58. Christopher T. Kello, Gordon D.A. Brown, Ramon Ferrer-i-Cancho, John G. Holden, Klaus Linkenkaer-Hansen, Theo Rhodes and Guy C. Van Orden. Scaling laws in cognitive sciences. – *Trends in Cognitive Sciences* 14(5) (2010) 223-232.
59. Y. Caron, P. Makris, N. Vincent. Use of power lawmodels in detecting region of interest. – *Pattern Recognition* 40 (2007) 2521-2529.
60. Eduardo G. Altmann, Janet B. Pierrehumbert, Adilson E. Motter. Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words. – *PLoS ONE* 4(11) (2009) e7678. – <www.plosone.org>, accessed 2011.11.19.
61. Luhn H.P. The automatic creation of literature abstracts. – *IBM Journal of Research and Development* 2 (1958) 159-165.
62. Robert M. Losee. Term Dependence: A Basis for Luhn and Zipf Models. – *Journal of the American Society for Information Science and Technology* 52(12) (2001) 1019-1025.
63. Janet I. Vousden. Units of English Spelling-to-Sound Mapping: A Rational Approach to Reading Instruction. – *Appl. Cognit. Psychol.* 22 (2008) 247–272.
64. Linda Alexander, Roger Johnson and John Weiss. Exploring Zipf's Law. – *Teaching Mathematics and its applications* 17(4) (1998).
65. Altmann G., v. Buttlar H., Rott W., Strauβ U. A law of change in language. – In: Brainerd B. (ed.) Historical linguistics. – Bochum: Brockmeyer, 1983. – P. 104-115.
66. Joshua B. Plotkin and Martin A. Nowak. Major Transitions in Language Evolution. – *Entropy* 3 (2001) 227–246.
67. Sebastian Bernhardsson, Luis Enrique Correa da Rocha, Petter Minnhagen. Size-dependent word frequencies and translational invariance of books. – *Physica A* 389 (2010) 330-341.
68. T. Maillart, D. Sornette, S. Spaeth, and G. von Krogh. Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution. – *Physical Review Letters* 101 (2008) 218701 .
69. Yaniv Dover. A short account of a connection of power laws to the information entropy. – *Physica A* 334 (2004) 591-599.
70. Jelena Nesterova. Spatial self-arrangement of expanding structures. 1. Overview of assessment concepts. – *Innovative Infotechnologies for Science, Business and Education* ISSN 2029-1035 – 2(9) (2010) 17-22.
71. Ramon Ferrer i Cancho. Decoding least effort and scaling in signal frequency distributions. – *Physica A* 345 (2005) 275–284.
72. R. Ferrer i Cancho. Zipf's law from a communicative phase transition. – *The European Physical Journal B* 47 (2005) 449–457.