ISSN 2224-087X. Електроніка та інформаційні технології. 2018. Випуск 9. С. 94–105
Electronics and information technologies. 2018. Issue 9. P. 94–105

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В НАУКОВИХ ДОСЛІДЖЕННЯХ

UDC 004.6, 004.9, 538.9

ZIPF'S AND HEAPS' LAWS FOR THE NATURAL AND SOME RELATED RANDOM TEXTS

O. Kushnir, V. Buryi, S. Grydzhan, L. Ivanitskyi, S. Rykhlyuk

Ivan Franko National University of Lviv
107 Tarnavsky Street, UA–79017 Lviv, Ukraine
o_kushnir@franko. lviv.ua

We have generated randomized Chomsky's texts and Miller's monkey random texts (RTs), basing on a source natural text (NT), and clarified their rank–frequency dependences, Pareto distributions, word-frequency probability distributions, and vocabularies as functions of text lengths. Here the Chomsky's RT is a NT randomized so that its 'words' represent any sequences of letters and blanks between the nearest occurrences of some preset letter (e.g., the letter *i*). We have compared the exponents appearing in different power laws that describe the word statistics for the NTs and RTs, and have analyzed how well theoretical relationships among those exponents are fulfilled in practice. We have proven empirically that the exponents α and β of the Zipf's law and the word probability distribution for the Chomsky's RTs are limited by the inequalities $\alpha < 1$ and $\beta > 1$, while their Heaps' exponent should be equal to $\eta \approx 1$. We have also compared our results to those obtained for the monkey texts. We have shown that the vocabulary of the Chomsky's texts is richer than that of the monkey texts. The Heaps' law is valid to extraordinarily good approximation for the Chomsky's RTs, similarly to the RTs generated by the intermittence silence process and unlike to sufficiently long NTs that reveal slightly convex vocabulary versus text length dependences plotted on the double logarithmic scale.

Key words: random texts, randomized texts, Miller's monkey texts, Chomsky's randomization, power laws, Zipf's law, Pareto distribution, word-frequency probability distribution, Heaps' law.

Introduction. Statistical regularities describing frequencies of occurrences of different linguistic elements in texts are widely studied in computational linguistics. They can provide the data important for information retrieval, intellectual data analysis, automated text indexing, and many other related fields [1]. Among statistical laws peculiar to individual texts and their corpora, Zipf's and Heaps' laws traditionally attract much attention of researchers [2]. The reasons lye both in their possible applications (e.g., in text categorization, stylometry, and language or author detecting [3]) and the studies of fundamental problems associated with either linguistic or purely statistical grounds for those rules [4–6]. Besides of usual natural texts (NTs), different kinds of randomized NTs and random texts (RTs) have often become subjects of extensive computational-linguistic researches [7–12]. In particular, this is caused by the efforts aimed at advanced distinguishing among the content-bearing (natural or artificial) messages and meaningless sequences of characters [13, 14].

Up to date, many types of artificial texts have been studied to some extent. In particular, these are well-known 'Miller's monkey' sequences (see [7]), RTs generated according to the Simon's model (see [2, 15]), and 'texts' obtained via different randomization procedures applied to NTs (see, e.g., [16]). An important idea lying behind the attention of researchers to the RTs is a potential feasibility of their numerical or even analytical analyses, using the probability theory. The main subject of the present work is statistical studies of the RTs generated using the algorithms close to the Chomsky's method, and comparison of the results with those derived for the initial NT.

Materials and methods. The NT subjected to our analysis was the J. R. R. Tolkien's novel "The Lord of the Rings" containing nearly 516 thousand words (see Table 1). We removed all of characters from the text except for letters (including those with diacritical signs usually present in the extended Latin alphabet) and did not discriminate among the lower- and upper-case letters. The compound hyphenated words were usually treated as single words and not split into elemental constituents. The contracted forms were reduced to their full equivalents (*Frodo's cracking* → *Frodo is cracking*, etc.), while the possessive nouns like *Frodo's* in *Frodo's fiftieth year* were left unchanged.

We generated two randomized versions of the original NT (abbreviated hereafter as NT0) according to the recipes very close to that suggested by N. Chomsky. They are denoted as RT1 and RT2. Additional RTs were also created, which were termed as RT3 and RT4 (see below). According to the basic Chomsky's method for randomizing NTs, the 'word' in a RT represents any sequence of 'letters' between the nearest occurrences of the letter *e*. The latter is the most frequent letter in English texts, so that its frequency is closest to that of the space as a word separator (see, e.g., [17]). From the data most often reported for the English language, the relative frequencies of the letter *e* and the space (*s*) are equal to $f_e = 0.125 \pm 0.005$ and $f_s = 0.18 \div 0.22$, respectively. The latter frequency implies that the average word length should be $l_{\text{av}} \approx 4.5$ letters, although one should remember that two alternative definitions, $f_s = N_s/(N_s + N_l)$ or $f_s = N_s / N_l$ (with N_l and N_s being the total amounts of letters and spaces, respectively), may be used in practice (see [11]). Notice that, for out text NT0, we had $f_s = 0.196 \div 0.244$ (l_{av} being from 4.1 to 5.1 letters), depending on the definition used.

Table 1

Some statistical characteristics of our NT and the related RTs

Text label	Total text length in letters (without blanks), 10^3	Total text length L in words, 10^3	Total vocabulary L, 10^3
NT0	2100.9	516.2	13.7
RT1	2482.0	134.7	94.8
RT2	1969.0	134.6	94.0
RT3	2100.9	412.9	18.7
RT4	333	50.0	8.5

Instead of *e*, in this work we used a different letter, *i* ($f_i = 0.070 \div 0.073$, with the letter-frequency rank ranging from $r = 5$ to $r = 7$ for different texts), so that the average 'word' length in the RTs was somewhat larger. Our first method (RT1) meant replacing the letters *i* with the spaces, while the spaces were replaced with the letters *i*. According to the second

method (RT2), we first removed the spaces from the text and then replaced the letters *i* with the spaces. The both methods were similar to the original Chomsky's recipe.

Our third method of producing RTs, RT3, meant removing the spaces from the text and then randomly generating them with a preset frequency, f_s (we chose, somewhat arbitrarily, the value 0.18). Notice that neither the total amount of letters nor the number of words in the NT is left unchanged by the randomizations procedures RT1 to RT3. We generated word separators in the text RT3 using a class Random(). To illustrate a diversity of RTs and take look at their possibly manifold statistical properties, we did not bother with a known problem of true randomness of the generator. Moreover, we deliberately tolerated its non-random biases originated, probably, from the use of current time as a parameter when seeding different instances of the class Random.

Finally, our last RT, RT4, was truly random rather than randomized, and so it had nothing to do with the initial NT. It corresponded to a family of 'monkey texts' that have been analyzed by B. Mandelbrot, H. Simon, G. Miller and N. Chomsky (see, e.g., [17]), with the word-separator frequency $f_s = 0.18$ and the equal frequencies of each letter $f_l = (1 - f_s)/M$, where M denotes the size of the 'alphabet'. We dealt with the two letters only ($M = 2$) and the whole text length was equal to $L = 5 \cdot 10^4$ words. No sequential chains of spaces were permitted in the RT4.

We calculated the absolute frequencies F of word types as functions of their rank r, the 'vocabulary' (i.e., the amount of different word types) V depending on the (variable) text length L, $V(L)$, the probability density $p(F)$ as a function of the frequency (which is often termed as a 'lexical frequency spectrum'), and the cumulative probability function $P(F)$ defined in a common manner as $P(F_0) = \Pr(F \geq F_0)$. Original software was developed for these purposes, using the language C#.

Results and their discussion. Fig. 1 displays the rank–frequency dependences calculated for all of our texts, NT0 and RT1–RT4. Usually, those dependences are treated as being described by a power law [4]:

$$F(r) \propto r^{-\alpha}. \tag{1}$$

It is known as a Zipf's law, with the exponent α being roughly equal to one. As seen from Fig. 1, all the dependences deviate to different extents from the straight lines on the double logarithmic scale, thus evidencing that the Zipf's law represents an approximation rather than a rigorous quantitative regularity. In particular, one observes a well-known departure from linearity of the $F(r)$ function for NT0 in the high-frequency region (for the ranks $r \leq 10 \div 20$), which is usually disregarded or partly mended with the Mandelbrot's correction, and a stepwise behaviour at the lowest frequencies, which is associated with massive sets of hapax legomena ($F = 1$), dis legomena ($F = 2$), etc. However, clear nonlinearities still persist in the middle-frequency region. In our opinion, the data is somewhat better described by a continuous, slowly varying increasing function $\alpha = \alpha(r)$ than by the idea of crossover between two power-law regions with different exponents, due to a transition from 'kernel' to 'unlimited' lexicons [18]. Note that the authors of Ref. [19] have called sufficiently long texts like our NT0 as 'saturated' texts, claiming a known phenomenon of 'convex' shape of their $\log F$ vs. $\log r$ curves (see also Ref. [20]). The slopes for the high-frequency ($20 < r < 500$) and low-frequency ($r > 500$) regions are $\alpha \approx 1.05$ and 1.58, respectively.

As evident from our empirical data (see Fig. 1 to Fig. 4), the RTs RT1 and RT2 reveal very similar statistical behaviours. Since our RTs are only single statistical realizations of the

randomization algorithms described in the previous section, one can hope that their properly averaged Zipf's curves, like all the other statistics, are the same. As a result, the randomization procedures abbreviated as RT1 and RT2 can be supposed to give the identical results. Like in the case of a similar RT studied in Ref. [17], RT1 and RT2 manifest a more pronounced linear $\log F(\log r)$ behaviour, the only exception being the lowest-rank region ($r < 20$). The Zipf's exponent α calculated outside the lowest-rank and staircase regions amounts to $\alpha_{RT1} \approx 0.91$.

Regarding the text RT3, it is unlike all the other texts in all respects (see the continuous lines in Fig. 1 to Fig. 4). The analysis shows that its characteristics are dominated by the specific features of practical work of the random generator utilized. In particular, all of the relevant empirical dependences are very far from the usual power law-like ones, with the only exception of the Heaps curve (see Fig. 4a and a further discussion).

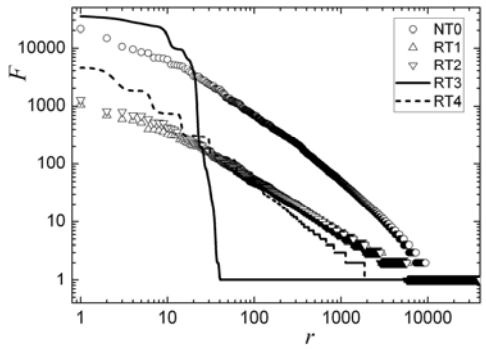

Fig. 1. Dependences of absolute word frequency F on the word rank r represented on log-log scale for the original NT (NT0) and the RTs (RT1, RT2, RT3 and RT4).

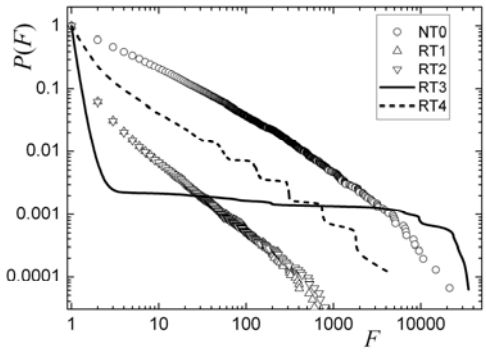

Fig. 2. Dependences of cumulative word probability P on the absolute word frequency F represented on log-log scale for the original NT (NT0) and the RTs (RT1, RT2, RT3 and RT4).

The 'intermittent silence process' underlying the text RT4 results in a staircase-like $F(r)$ dependence, which is partly observed even at the lowest ranks. The reason is easily understood: all the 'words' having the same lengths (1, 2, ... letters) have equal probabilities and so equal frequencies, whereas the incomplete staircase behaviour for low ranks is due to insufficient statistics, i.e. due to finite size of the text. The linear fitting for RT4 performed outside the regions of the lowest and highest ranks yields in the Zipf's constant α equal to 1.20.

Fig. 2 presents the cumulative probability distribution for the texts NT0 and RT1–RT4, or a so-called Pareto function. Since the dependence $P(F)$ is in fact a renormalized inverse function of $F(r)$, one gets (see, e.g., [4, 21])

$$P(F) \propto F^{-\pi}, \tag{2}$$

$$\pi = \alpha^{-1}. \tag{3}$$

Eqs. (2) and (3), however, do not consider a fine though important point: because of its staircase nature, the dependence $F(r)$ cannot be plainly inverted into $P(F)$. To do this, one has to get rid of the stairs, through assigning the same rank to different word types having equal frequencies. Owing to this peculiarity, which is generally neglected in the literature, the link between formulae (1)–(3), and so the relation between the exponents α and π, are not so straightforward.

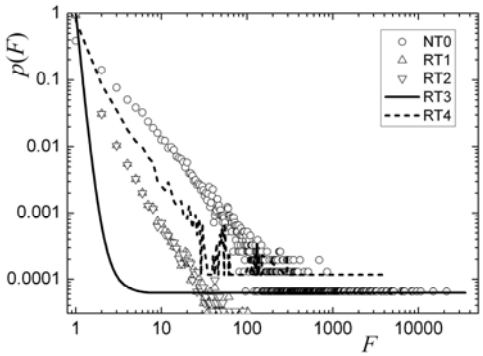

Fig. 3. Dependences of mass probability function p on the absolute word frequency F represented on log-log scale for the original NT (NT0) and the RTs (RT1, RT2, RT3 and RT4).

As seen from Fig. 2, the $P(F)$ curves deviate from linearity, especially in the regions of the lowest and highest frequencies. The slopes for the texts NT0 and RT1 (or RT2) estimated in the intermediate region are respectively $\pi_{NT0} \approx 0.87$ and $\pi_{RT1} \approx 1.15$, thus giving $\alpha_{NT0} \approx 1.15$ and $\alpha_{RT1} \approx 0.87$. While it is difficult to assign a single-valued Zipf's exponent to NT0 (see above), the latter α_{RT1} value agrees satisfactorily with that obtained from the data of Fig. 1.

Performing the same procedure for RT4, one obtains a rough estimation $\pi_{RT4} \approx 0.86$. The value $\alpha_{RT4} \approx 1.16$ calculated on this basis with Eq. (2) correlates moderately with the direct result $\alpha_{RT4} \approx 1.20$. Notice also that the effect of quasi-stepwise $P(F)$ behaviour for the 'monkey texts' has earlier been revealed by Bernhardsson et al. [12]. These authors believe that the 'true' Pareto index π corresponding to the smooth theoretical function (2) is a slope of the straight line that corresponds to the envelope of the actual $\log P(\log F)$ dependence.

The probability density functions for out texts are depicted in Fig. 3. Disregarding, as before, a specific $p(F)$ dependence for RT3, one can notice that all the other curves are qualitatively similar to those typical for the NTs. It is well-known (see, e.g., Ref. [4]) that the $p(F)$ function should behave according to the power law:

$$p(F) \propto F^{-\beta}, \tag{4}$$

with the constant index β linked to the Zipf's exponent via

$$\beta = 1 + 1/\alpha = 1 + \pi. \tag{5}$$

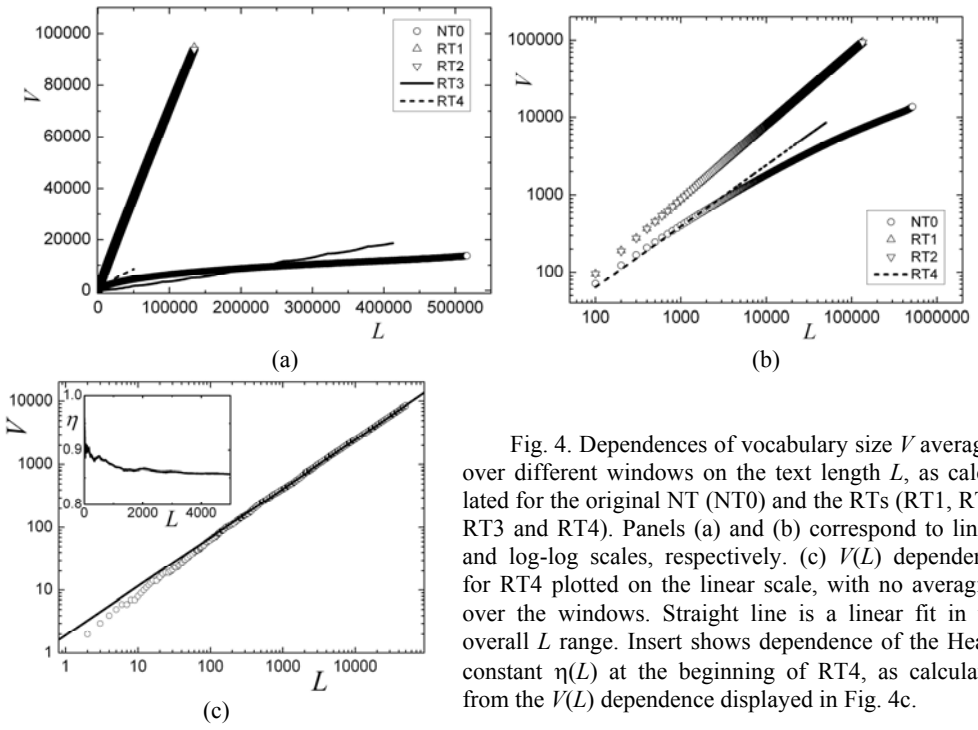

Fig. 4. Dependences of vocabulary size V averaged over different windows on the text length L, as calculated for the original NT (NT0) and the RTs (RT1, RT2, RT3 and RT4). Panels (a) and (b) correspond to linear and log-log scales, respectively. (c) $V(L)$ dependence for RT4 plotted on the linear scale, with no averaging over the windows. Straight line is a linear fit in the overall L range. Insert shows dependence of the Heaps constant $\eta(L)$ at the beginning of RT4, as calculated from the $V(L)$ dependence displayed in Fig. 4c.

In general, analyzing quantitatively the dependences of Fig. 3 and deriving β on this basis have some limitations [4, 22–26]. Being in fact a derivative of the Pareto function, the $p(F)$ function manifests much more noise. This is readily confirmed by linear fitting of the $p(F)$ curve for the text NT0 plotted on the log-log scale. Even after excluding from consideration the most noisy distribution 'tail', we obtain the exponent $\beta_{NT0} \approx 1.56$ which, according to formula (5), leads to a completely unreliable result, $\alpha_{NT0} \approx 1.79$. Indeed, the high-rank region is roughly characterized by the value 1.58, not to mention a still less α_{NT0} peculiar of the lower ranks. The results become still worse if the lowest-frequency data (in particular, the first 11 data points in Fig. 3) are used, although they embrace a great bulk of the word types and are often used in the fitting (see [23]). Then we get $\beta_{NT0} \approx 1.44$ (and so $\alpha_{NT0} \approx 2.27$), which is far from the real Zipf's exponents. In this respect, it would be better to derive the β exponent using the analytical techniques like those presented in Refs. [4, 22–26], instead of linear fitting.

Surprisingly, the results greatly improve for the case of RT1 or RT2 and, moreover, the appropriate curves are less noisy and closer to linearity. Discarding the first point of the probability distribution (i.e., the region with the steepest slope) and its noisy tail, we obtain $\beta_{RT1} \approx 2.15$. With Eq. (5), this implies $\alpha_{RT1} \approx 0.87$, in a good agreement with the values 0.91 and 0.87 derived following from the data of Fig. 1 and Fig. 2.

The 'spikes' observed in the data generated by the intermittence silence process (see the dashed line in Fig. 3 for RT4) are already well-known [8, 10, 12]. The linear fitting on the log-

log scale results in the exponent $\beta_{RT4} \approx 1.89$ (i.e., $\alpha_{RT4} \approx 1.12$), which fairly agrees with the values 1.83 (1.20) and 1.86 (1.16) obtained using respectively the Zipf's and Pareto dependences. According to the theory, the β exponent is given by the general formula [12]

$$\beta = \frac{2\ln M - \ln(1-f_s)}{\ln M - \ln(1-f_s)}. \quad (6)$$

In our case ($M = 2$ and $f_s = 0.18$) Eq. (6) yields the value $\beta_{RT4} \approx 1.78$, which is not so far from our estimations for RT4. Some discrepancy between the theory and the empirical data can originate from the fitting itself, as well as from relatively short length of the text and the accompanying finite-size effects.

Fig. 4 shows the $V(L)$ functions for our NT and RTs on the linear and log-log scales. Here the exception is the vocabulary growth curve for the text RT3, which acquires irregular shape on the double logarithmic scale and so has been omitted in Fig. 4b. Note that, for eliminating the noise, the $V(L)$ dependences presented in Fig. 4a, b have been averaged over moving windows of the lengths L, with the minimal window size and the window shift step equal to 100 words. For the sake of comparison, we also present in Fig. 4c a non-averaged $V(L)$ dependence for RT4. It reveals some fluctuations, which are the best observed at small L. Finally, the total vocabulary sizes for all the texts under test are collected in Table 1.

Although some authors suggest complicated theoretical functions for the dependence of the vocabulary on the text size (see, e.g., Refs. [20, 27]), this dependence is commonly represented by a power Heaps' law [28–30],

$$V(L) \propto L^{\eta}. \quad (7)$$

Here the constant η, or the Heaps exponent, is linked with the other parameters as follows (see, e.g., Ref. [31]):

$$\eta = \beta - 1 = 1/\alpha \ (\alpha \geq 1),$$
$$\eta = 1 \ (\alpha < 1). \quad (8)$$

The validity of formula (7) is clearly evidenced by the log-log plots of Fig. 4b for the texts RT1 and RT2, where the slopes are $\eta_{RT1} \approx 0.95$ and $\eta_{RT2} \approx 0.96$ and the coefficients of determination $R^2 \approx 0.99999$. A close proximity of the Heaps exponents to one is also confirmed by the $V(L)$ functions plotted on the linear scale. Here the linear fits, which are not shown in Fig. 4a, are also satisfactory ($R^2 \approx 0.999$). As seen from the above discussion, the texts RT1 and RT2 are characterized by the Zipf's exponent clearly less than one. According to formulae (8), then the Heaps exponent has to be exactly one, i.e. the vocabulary has to increase linearly with increasing text size. Our empirical data in fact prove this feature of the RTs generated along the Chomsky's method. We believe that the small discrepancy between $\eta_{RT1,2}$ and the unit value should originate from the calculation inaccuracies, finite-size effects and the limitations associated with a single statistical realization of the randomization procedure. To remove the latter limitation, one must generate a sufficiently large sample (say, 10^3 or 10^4) of the RTs and calculate their statistically averaged parameters. Finally, we are to notice that the same statistical feature, $\eta \sim 1$, is typical for the RTs generated by the simplest version of the Simon's growth model (see Refs. [2, 15]).

In general, the Heaps' law fulfils well for RT4. Using the averaged data of Fig. 4b, one obtains $\eta_{RT4} \approx 0.79$ as a slope and a very high R^2 (larger than 0.999999). These results confirm the conclusion [12] that the monkey books obey the Heaps' law extremely well. Following

from the β value derived earlier (1.78) and formulae (8), we get the theoretical η exponent equal to 0.78, which agrees with our empirical data within the limits of the fitting errors.

Notice also that, according to formula (6), the Heaps exponent for the monkey text cannot reach one, $\eta_{RT4} < 1$, irrespective of the M and f_s values. This is an unobvious property because a naive reasoning may have assumed just the contrary: the vocabulary of the text where any combinations of symbols are permitted should seemingly grow much more rapidly, say, linearly with increasing text size. On the other hand, the same intuitive considerations concerning the vocabulary of the Chomsky's RTs would have led to the conclusions about slower vocabulary growth ($\eta_{RT1,2} < 1$). Indeed, it is evident that NTs reveal much poorer vocabularies than RTs, of which vocabularies are limited only by a number of possible combinations of letters. The Chomsky's RTs are randomized NTs not totally random and, therefore, at least some minimal portion of the initial ('true') words can, in principle, survive the randomization process, thus not contributing to rapid vocabulary increase, as compared with purely RTs. This again contradicts our empirical data, though it would be instructive to substantiate the fact $\eta_{RT1,2} \approx 1$ basing on thorough theoretical grounds.

Some attention should be paid to the smallest-L region of the $V(L)$ function for RT4. Issuing from purely computational reasons, it would be convenient to analyze this region, using a 'raw', non-averaged dependence $V(L)$. Since the latter is not being averaged over different window positions in the text, we have allowed it to contain much more detailed data, including in the region under test (see Fig. 4c). This is unlike the averaged $V(L)$ curve in Fig. 4b where the small-L region is poorly represented. The linear fit performed in the overall abscissa range in Fig. 4c results in the slope $\eta_{RT4} \approx 0.78$, thus agreeing perfectly with the theory. Nonetheless, the initial part of the $V(L)$ dependence persists in deviating from the linear trend that dominates elsewhere. The most convenient mathematical way for expressing this phenomenon would be assuming that $\eta = \eta(L)$. Fig. 4c, insert, shows the Heaps exponent obtained as a (non-smoothed) logarithmic derivative in the initial ($L < 5000$) part of the $V(L)$ dependence. It testifies a presence of a 'transition process' in the vocabulary growth, which is accompanied with a rapid decrease in the η exponent (from ~ 1.0 at $L = 1$ to about 0.85 already at $L = 4000$). In spite of its small importance under the conditions of infinite text length increase, the transition process mentioned represents a principled empirical fact available for any text characterized with $\eta < 1$. It emphasizes that the theoretical curve given by Eq. (7) cannot describe the data at $L \sim 1$ where we have $V \sim L$.

As seen from Fig. 4b, the NT, NT0, demonstrates the most obvious departure from the Heaps' law. As with RT4, this phenomenon can be treated using a slowly varying function of the text length, $\eta = \eta(L)$. In the regions of small and large L, we obtain respectively $\eta_{NT0} \approx 0.68$ and $\eta_{NT0} \approx 0.48$ ($R^2 \approx 0.999$). Hence, one can conclude that the convexity of the $V(L)$ dependence plotted on the log-log scale represents a notable feature of moderately long NTs and their important difference from the RTs generated using the intermittence silence and the Chomsky's algorithm.

Finally, the $V(L)$ function for RT3 (see Fig. 4a) offers a curious example of possible vocabulary growth curves peculiar for artificial RTs. While the overall tendency is a linear relation between V and L, the curve also manifests a regular, almost periodic structure, with the period roughly equal to $L_T \sim 54000$. It seems reasonable to assume that periodic 'accelerations' and 'decelerations' of the vocabulary growth are associated with specific features of the random generator employed to arrange the word separators. Instead of being truly 'random', the

latter arranges the blanks so as to first generate long (and so mostly nonrecurring) 'words', and then produce a great number of short 'words', which naturally appear to be the same more frequently. This implies that the text-developing process includes alternating stages of faster and slower vocabulary growths, thus resulting in quasi-periodicity of the Heaps curve $V(L)$.

Conclusions. We have generated a number of randomized texts based on a source NT (NT0) and have clarified statistical regularities of their lexical sets. Our RTs have been produced using the algorithms close to that suggested by N. Chomsky (RT1 and RT2), as well as by the 'intermittence silence' algorithm (RT4). To obtain a broader scale of RTs, we have also employed a random-like recipe for word separating in the initial text (RT3). Among different statistical characteristics, we have studied the rank–frequency dependence, the Pareto distribution, the lexical frequency spectrum, and the vocabulary as a function of the text length.

The main findings of the present work can be summarized as follows. First, we have demonstrated that relatively long ($5 \cdot 10^5$ words or longer) NTs manifest apparently more convex Zipf's curves than the Chomsky's RTs, of which rank–frequency dependences are approximately linear on the log-log scale. The latter is also peculiar for the envelope of the Zipf's curve for the monkey text, in spite of quasi-staircase behaviour of the latter curve.

We have elucidated the problem of deriving the exponents appearing in different power laws that describe the word statistics of the NTs and RTs, and have analyzed to which extent the theoretical relationships among those exponents are fulfilled in practice. In particular, the lexical-frequency spectral function $p(F)$ for the Chomsky's text reveal less fluctuations than that for the NT and, therefore, the exponents α and β found respectively from the $F(r)$ and $p(F)$ dependences correlate better. We have also proven empirically that the word-statistics exponents α *and* β for the Chomsky's texts are limited by the inequalities $\alpha < 1$ and $\beta > 1$. Then the Heaps exponent for this type of RTs should be equal to $\eta \approx 1$, in agreement with our data. This situation is similar to the RTs produced in frame of the simplest Simon's model. We have also demonstrated that the inequality $\eta_{RT1} > \eta_{RT4}$ is valid for the Heaps exponents of the Chomsky's and Miller's monkey texts. Notice that the power-law exponent β for our monkey text (with the alphabet size $M = 2$) agrees with the established theoretical value.

We have confirmed empirically that the Heaps exponent is less than one for the 'monkey text' RT4 and found that it is very close to one for the texts RT1 and RT2 randomized according to N. Chomsky. One can reformulate these facts as a following counter-intuitive statement: the vocabulary of the randomized Chomsky's texts is richer than that of the monkey texts. It is important that the statistical properties of the Chomsky-like texts RT1 and RT2 are in fact identical and are not affected by the differences of practical algorithms used for their generation.

We have empirically confirmed the statement of Ref. [12] that the Heaps' law is valid to extraordinarily good approximation for the monkey texts and have demonstrated for the first time that the same is true for the randomized Chomsky's texts. This is in a clear contrast with relatively long NTs, which reveal slightly convex vocabulary–text length dependences plotted on the double logarithmic scale.

Concerning the problem of distinguishing among the NTs and RTs, the artificial monkey texts can be easily recognized by their spike-like lexical-spectrum dependences, whereas the artificial Chomsky's texts can be identified by a linear growth of their vocabulary on the text length. Differentiation of the latter texts from the Simon's ones represents a separate problem.

The authors thank Ph. D. Student, Assist. Prof. Kushnir L. O. for providing a carefully read copy of ASCII-coded Tolkien's novel.

REFERENCES

1. *Manning C. D.* Foundations of statistical natural language processing / Manning C. D., Schutze H. – London: The MIT Press Cambridge, 1999. – 680 p.
2. *Zanette D. H.* Statistical patterns in written language / Zanette D. H. – Centro Atomico Bariloche, 2012, 87 p. http://fisica.cab.cnea.gov.ar/estadistica/2te/
3. *Damashek M.* Gauging similarity with n-grams: language-independent categorization of text / Damashek M. // Science. – 1995. – Vol. 267. – P. 843–848.
4. *Newman M. E. J.* Power laws, Pareto distributions and Zipf's law / Newman M. E. J. // Contemporary Phys. – 2005. – Vol. 46. – P. 323–351.
5. *Ferrer i Cancho R.* Least effort and the origins of scaling in human language / Ferrer i Cancho R., Sole R. V. // Proc. Nat. Acad. Sci. – 2003. – Vol. 100. – P. 788–791.
6. *Manin D. Y.* Zipf's law and avoidance of excessive synonymy // Cognitive Sci. – 2008. – Vol. 32. – P. 1075–1098.
7. *Li W.* Random texts exhibit Zipf's-law-like word frequency distribution / Manin D. Y. // IEEE Trans. Inform. Theory. – 1992. – Vol. 38. – P. 1842–1845.
8. *Ferrer i Cancho R.* Zipf's law and random texts / Ferrer i Cancho R., Sole R. V. // Adv. Complex Syst. – 2002. – Vol. 5. – P. 1–6.
9. *Krause A.* Not so randomly typing monkeys – Rank-frequency behavior of natural and artificial languages / Krause A., Zollmann A. // Algorithms for Information Networks – Project Report. 2005. http://www.cs.cmu.edu/~zollmann/publications.html
10. *Biemann C.* A random text model for the generation of statistical language invariants / Biemann C. // Int. Conf. "Word Sense Induction and Disambiguation at Powerset". – Microsoft Research, Redmond, WA, USA, Sept. 2008. – P. 1–8.
11. *Ferrer-i-Cancho R.* Random texts do not exhibit the real Zipf's law-like rank distribution / Ferrer-i-Cancho R., Elvevåg B // PLoS ONE. – 2010. – Vol. 5. – e9411 (10 p.).
12. *Bernhardsson S.* A paradoxical property of the monkey book / Bernhardsson S., Baek S. K., Minnhagen P. // J. Statist. Mech.: Theory and Exper. – 2011. – P07013 (12 p.).
13. *Cohen A.* Numerical analysis of word frequencies in artificial and natural language texts / Cohen A., Mantegna R. N., Havlin S. // Fractals. – 1997. – Vol. 5. – P. 95–104.
14. *Casti J. L.* Bell curves and monkey languages // Complexity. – 1995. – Vol. 1. – P. 12–15.
15. *Zanette D. H.* Dynamics of text generation with realistic Zipf's distribution / Zanette D. H., Montemurro M. A. // J. Quant. Linguist. – 2005. – Vol. 12. – P. 29–40.
16. *Montemurro M. A.* Long-range fractal correlations in literary corpora / Montemurro M. A., Pury P. A. // Fractals. – 2002. – Vol. 10. – P. 451–461.
17. *Young C.* Who's afraid of George Kingsley Zipf // Significance. – 2013. – Vol. 10. – P. 29–34.
18. *Ferrer i Cancho R.* Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited / Ferrer i Cancho R., Solé R. V. // J. Quant. Linguist. – 2001. – Vol. 8. – P. 165–173.
19. *Пиотровский Р. Г.* Статистические модели текста и опыт их лингво-синергетического анализа / Пиотровский Р. Г. // Научно-техн. информ., сер. 2. – 2007. – №8. – С. 1–11.

20. *Font-Clos F.* Log-log convexity of type-token growth in Zipf's systems / Font-Clos F., Corral A. // Phys. Rev. Lett. – 2015. – Vol. 114. – 238701 (4 p.).
21. *Li W.* Fitting ranked linguistic data with two-parameter functions / Li W., Miramontes P., Cocho G. // Entropy. – 2010. – Vol. 12. – P. 1743–1764.
22. *Adamic L. A.* Zipf, power-laws, and Pareto – a ranking tutorial // Xerox Palo Alto Research Center. – 2000. http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html
23. *Goldstein M. L.* Problems with fitting to the power-law distribution / Goldstein M. L., Morris S. A., Yen G. G. // Eur. Phys. J. B. – 2004. – Vol. 41. – P. 255–258.
24. *Clauset A.* Power-law distributions in empirical data / Clauset A., Shalizi C. R., Newman M. E. J. // SIAM Rev. – 2009. – Vol. 51. – P. 661–703.
25. *Corral A.* A practical recipe to fit discrete power-law distributions / Corral A., Deluca A., Ferrer-i-Cancho R. – 2012. arXiv:1209.1270 [stat.AP]
26. *Corral A.* Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions / Corral A., Deluca A. // Acta Geophys. – 2013. – Vol. 61. – P. 1351–1394.
27. *Bochkarev V. V.*. Deviations in the Zipf and Heaps laws in natural languages / Bochkarev V. V., Lerner E. Yu., Shevlyakova A. V. // J. Phys.: Conf. Ser. – 2014. – Vol. 490. – 012009 (4 p.)
28. *van Leijenhorst D. C.* A formal derivation of Heaps' law / van Leijenhorst D. C., van der Weide Th. P. // Inform. Sci. – 2005. – Vol. 170. – P. 263–272.
29. *Gerlach M.* Stochastic model for the vocabulary growth in natural languages / *Gerlach M., Altmann E. G. //* Phys. Rev. X. – 2013. – Vol. 3. – 021006 (10 p.).
30. *Font-Clos F.* A scaling law beyond Zipf's law and its relation to Heaps' law / Font-Clos F., Boleda G., Corral A. *// New* J. Phys. – 2013. – Vol. 15. – 093033 (16 p.).
31. *Sano Y.* Zipf's law and Heaps' law can predict the size of potential words / Sano Y., Takayasu H., Takayasu M. // Progr. Theor. Phys., Suppl. – 2012. – No. 194. – P. 202–209.

Стаття: надійшла до редакції 12.12.2017,
доопрацьована 18.12.2017,
прийнята до друку 20.12.2017.

ЗАКОНИ ЦІПФА І ГІПСА ДЛЯ ПРИРОДНОГО ТЕКСТУ ТА ДЕЯКИХ РАНДОМНИХ ТЕКСТІВ НА ЙОГО ОСНОВІ

О. Кушнір, В. Бурий, С. Гриджан, Л. Іваніцький, С. Рихлюк

Львівський національний університет імені Івана Франка
вул. Ген.Тарнавського, 107, 79017 Львів, Україна
o_kushnir@franko. lviv.ua

На основі вихідного природного тексту згенеровано рандомізовані тексти Хомського і рандомні тексти "мавпи Міллера". Рандомні тексти створено за таким алгоритмом: усі літери мають однакову наперед задану ймовірність, а ймовірність розділювача поміж словами (пробілу) задається незалежно від них. Вивчено залежності ранг–частота, розподіли кумулятивної ймовірності Парето, розподіли ймовірності частоти слів і залежності кількості різних слів (словники) від кількості всіх слів як функції довжини тексту. Під рандомними текстами Хомського розуміємо природний текст, рандомізований так, що "слова" в ньому

є довільними послідовностями літер і пробілів між найближчими появами деякої наперед визначеної літери (наприклад, *і*). Виконано порівняння показників степенів, які фігурують у різних степеневих законах, що описують статистику слів для природного тексту і рандомного тексту, а також проаналізовано, наскільки теоретичні співвідношення між цими степенями дотримано на практиці. Згадані співвідношення дещо нагадують аналоги так званих співвідношень універсальності поміж степенями різних критичних параметрів у фізиці критичних явищ. Емпірично доведено, що показники α і β законів Ціпфа і розподілу ймовірності слів для рандомних текстів Хомського обмежені нерівностями $\alpha < 1$ і $\beta > 1$, тоді як показник закону Гіпса для словника повинен становити $\eta \approx 1$. Ці результати порівняно з даними для текстів мавпи Міллера. З'ясовано, що словник текстів Хомського багатший, ніж словник текстів мавпи Міллера. Виявлено, що закон Гіпса для рандомних текстів Хомського виконується з винятковою точністю, схоже до рандомних текстів, генерованих згідно з процесом "intermittence silence". Це дещо відмінне від ситуації для достатньо довгих природних текстів, які виявляють дещо "випуклу" залежність словника від довжини тексту, побудовану в подвійному логарифмічному масштабі.

Ключові слова: рандомні тексти, рандомізовані тексти, тексти мавпи Міллера, рандомізація Хомського, степеневі закони, закон Ціпфа, розподіл Парето, розподіл імовірності частоти слів, закон Гіпса.