UDC 004.6, 004.9, 538.9

RANK DEPENDENCES AND LEXICAL FREQUENCY SPECTRA FOR THE SUBGROUPS OF DIFFERENT-LENGTH WORDS IN TEXTS

O. S. Kushnir, M. Ya. Maksysko, L. B. Ivanitskyi, S. V. Rykhlyuk

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

In this work we have found statistical word-length distribution for the text of J. R. R. Tolkien's novel "The Lord of the Rings". We have studied 'partial' dependences rank–frequency $f(r)$ and probability density functions $p(f)$ for the subgroups of words with different lengths ($l = 1$–16 letters), as well as the corresponding 'combined' dependences for the words of all lengths. It has been revealed that the "partial" dependences are in general worse described by the power laws, known as Zipf's ones, than the 'combined' ones. The functions $f(r)$ and $p(f)$ for the intermediate lengths $l = 5$–10 are the closest to the power law, with the exponents $\alpha \approx 1$ for $f(r)$ and $\beta \approx 2$ for $p(f)$. The hypothesis has been put forward that the rank dependences for the smallest lengths can turn out to be close to exponential $(\alpha \to \infty)$, similar to the $f(r)$ function for some Eastern languages with a limited vocabulary, while the rank dependence and the probability density for the largest l can tend respectively to logarithmic $(\alpha \to 0)$ and exponential $(\beta \to \infty)$ functions.

Key words: statistical distributions, power distributions, rank dependences, computational linguistics, natural languages, Zipf's law, word-length distribution.

Introduction. Rank–frequency dependences and spectral frequency distributions for the words in natural and artificial texts represent a traditional subject of studies in the fields of computational and statistical linguistics. Recently there has been an upsurge in the interest of researchers to the problems of fulfilment and specific features of those dependences concerned with various linguistic subsystems rather than the whole system of words. In particular, this can be the sequences of letters (so-called 'letter N-grams' with different N – beginning from N = 1, i.e. the case of separate letters, and ending with N as high as, e.g., 15) and the sequences of words, or 'word N-grams' (including relatively long 'phrases', e.g., those characterized by N = 5). These studies have been performed for a numbers of languages with various topologies (see, e. g, Refs. [1–3]). They are useful in understanding fundamental reasons for the universal power laws holding true in statistical linguistics.

Besides of division of a text into subsystems of N-grams with different lengths N, there can be many other divisions into subsystems, for example into subsets of words having different lengths defined in the units of number of letters (see [4]). The appropriate results seems to be important at least because the analysis of the corresponding subsets for the cases of random or randomized analogues of the natural texts can often allow for deriving more or less exact numerical, or even analytical, solutions. Then the natural and random texts can be compared to one another and their mutual and opposite features can be elucidated.

The aim of the present work is studying the main peculiarities of the power laws known as Zipf's ones for the lexical subsystems of a natural text embracing the words of different lengths.

Materials and methods. We studied the text of J. R. R. Tolkien's novel "The Lord of the Rings" pre-processed in a usual manner and written down in ASCII codes. It included about 516,000 word tokens. Using the environment 'Visual Studio 2012', we developed a program in the language C#. It calculated rank–frequency dependences, $f(r)$, and word frequency spectra, $p(f)$. The former are dependences of the relative frequency f of word types ($f = F/L$, with F being the absolute frequency, i.e. the number of occurrences of a given word token in a text, and L implying the total number of word tokens in that text, i.e. the text length) on their rank r (i.e., a sequential number of a word type in the frequency list arranged in descending order). The latter represent frequency dependences of the probability p of a word type ($p = N/V$, with N and V denoting respectively the number of word types having the frequency F and the total number of different word types) on the frequency f.

Additionally, we determined statistical distribution of words, $p(l)$, depending on their length l. The probability distribution obtained for our text is displayed in Fig. 1. The most probable word length is approximately $l \approx 4$, i.e. somewhat less than that widely known for the English language ($l \approx 5$). Probably, this can be explained as a specific feature of Tolkien's writing. Notice that the data $p(l)$ can be approximately described by the log-normal distribution. Following from the quantitative behaviour of our $p(l)$ function, we restricted all the subsequent statistical calculations to the length interval $l = 1$–16. Moreover, the statistics for some of the smallest and largest l's were also insufficient. For this reason, the $f(r)$ dependences for the cases $l = 1$ and 2 were not interpreted quantitatively.

Results and their discussion. Fig. 2a illustrates some examples of the rank dependences $f(r)$ (at $l = 4$ and 11), and a 'combined' rank–frequency dependence obtained for all of the words ($l = 1$–25). Fig. 2b shows the spectral dependences calculated for $l = 5$ and 10 as examples, together with a 'combined' $p(f)$ dependence.

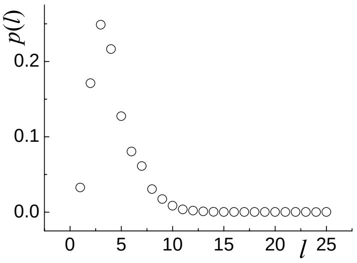


Fig. 1. Probability p of occurrence, in the text under study, of the words having the lengths l defined in the number of letters.

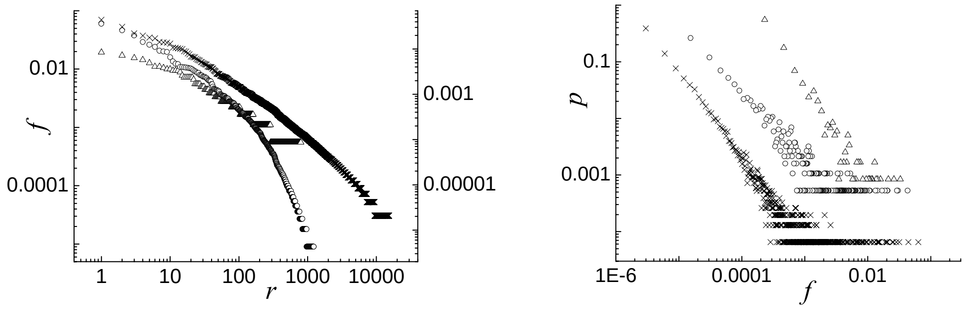
According to the theory, the dependence $f(r)$ should be described, at least approximately, by the relation referred sometimes to as a first Zipf's law (see, e.g., [5]):

$$f(r) \propto r^{-\alpha}, \tag{1}$$

where a constant α is called a Zipf's exponent. The *p*(*f*) dependence is given by a so-called second Zipf's law [5]:

$$p(f) \propto f^{-\beta}, \tag{2}$$

with the exponent β being also a constant. Notice that the both formulae (1) and (2) correspond to the power laws, which is a reason why the double logarithmic scales are used in Fig. 2.



(a) (b)

Fig. 2. (a) Rank–frequency dependences *f*(*r*) for the words with the lengths *l* = 4 (circles) and *l* = 11 (triangles), and 'combined' dependence *f*(*r*) for all the words (crosses). (b) Frequency–probability dependences *p*(*f*) for the words with the lengths *l* = 5 (circles) and *l* = 10 (triangles), and 'combined' dependence *f*(*r*) for all the words (crosses).

As a matter of fact, most of the 'partial' dependences *f*(*r*) and *p*(*f*) for individual lengths *l* can be treated in terms of the power law, though with some caution. It is known that, for large enough texts or corpora, both the rank–frequency and the frequency–probability functions for all the words (i.e., the 'combined' dependences, in terms of the present work) reveal a power law as a general tendency only, not a thorough mathematical rule. As a consequence, the exponents α are usually derived with a standard graphical method only for the central regions, or 'cores' of the $\log f$ vs. $\log r$ dependences, where they are approximately linear. Deviations from the power law turn out to be still larger for the dependences *f*(*r*) and *p*(*f*) obtained for the individual lengths *l*.

Nonetheless, we have used the same technical approaches as for the 'combined' $\log f$ vs. $\log r$ dependences, and found the α's as slopes. This has been done mainly for the ranks $r = 20$–500, if the statistics is sufficient. One has to be aware of some ambiguity of our quantitative α data. For instance, for the case of $l = 3$ we obtain $\alpha \approx 2.75$ (see Fig. 3) in the more or less 'standard' (see above) rank interval ranging from $r = 20$ to $r = r_{\max} = 390$. However, evident nonlinearities in the $\log f$ vs. $\log r$ dependence persist as far as up to $r \sim 40$, whereas a non-negligible contribution to the 'stairs' observed in the end of the frequency list originates from such untypical 'words' as, e.g., *cht* (having happened in the Language Appendix of the Tolkien's novel). Then further contraction word type list and the rank range to $r = 40$–310 leads to somewhat different result, $\alpha \approx 2.91$, with a notably higher goodness of fitting (the coefficient of determination being $R^2 = 0.997$, rather that 0.991, as in the case of the wider rank range). The final results for the α (l) dependence are shown in Fig. 3.

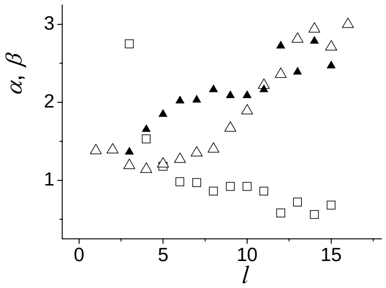


Fig. 3. Dependences of exponents α (squares) i β (triangles), which appear in the first and second Zipf's laws (see formulae (1) and (2)), as determined empirically for the word subsets with the lengths l = 1–16. Black triangles correspond to the β values calculated using formula (5) and the more reliable empirical data α.

It is interesting that the function $f(r)$ at l = 2 is poorly described by a power law (cf. with the cases of l = 4, 11 in Fig. 2a). Instead, it is closer to either logarithmic or exponential functions (see Fig. 4). One cannot exclude that both of them have some grounds. Indeed, the logarithmic function of the form

$$f(r) \propto \log(r^{-1}) \tag{3}$$

is known to govern, at least qualitatively, the rank–frequency dependences for the letters [6, 7] and, moreover, there can exist some analogies between the letters as linguistic elements and the shortest words with l = 1 or 2. Nonetheless, these analogies do not consider a drastic difference between the graphemes or phonemes (e.g., written down as *i* and *th*) that lack any semantics – and one- or two-letter words like *I* and *be* that bear clear semantic constituents.

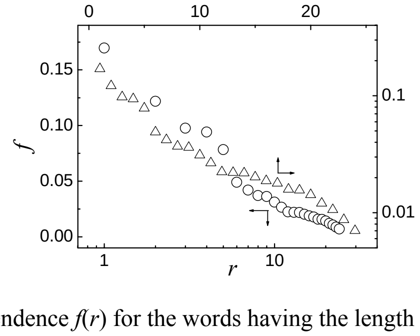


Fig. 4. Rank–frequency dependence $f(r)$ for the words having the length l = 2 plotted in the coordinates f vs. $\log r$ (circles) and $\log f$ vs. r (triangles), which try to verify the accordance with the theoretical dependences given by formulae (3) and (4).

On the other hand, the list of the words with $l = 1$ or 2 is very limited, thus putting forward a reasonable analogy with the $f(r)$ functions typical for the words originated from the languages with a limited vocabulary (e.g., Chinese, Korean and Japanese – see Ref. [8]). It is known that the rank–frequency dependences for those languages manifest an exponential 'regime' [8]:

$$f(r) \propto \exp(-ar)\,, \tag{4}$$

where a denotes a constant. Curiously enough, one cannot give preference to one of the functions (3) or (4) with our statistical data, since the data set is not large enough (see Fig. 4). In other words, much larger texts, or even corpora, are needed to resolve this controversy.

Beginning from the length $l = 3$, the power-law hypothesis for description of the rank–frequency dependences becomes dominating and neither logarithmic nor exponential function can rival the model given by formula (1), in spite of the quantitative difficulties of the latter already mentioned above.

The dependence $\beta(l)$, which has been derived basing upon our $\log p$ vs. $\log f$ data like those illustrated in Fig. 2b, is also shown in Fig. 3. There are considerable problems with accurate determination of the β index using simple graphical methods (see, e.g., Refs. [5, 9, 10] for details). The relevant data almost always reveal large calculation errors, even larger that those peculiar for the $\alpha(l)$ data. This comes from conspicuous noises contained in the $\log p(\log f)$ dependence in any practical situation. As a practical means, we have calculated the β exponents using a linear fit of the first 10–30 frequency points of the $p(f)$ dependences, which involve an overwhelming majority of the word tokens present in the text.

It is useful to check a consistency of independently calculated α and β parameters using a known theoretical relationship (see, e.g., [5, 11]),

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

Here we should admit poor correlation between the exponents α and β calculated for the case of 'combined' word lengths, which is the most abundant in the statistically sense. Indeed, we obtain $\alpha \approx 1.05$ for a central 'core' ($r = 20$–500) of the 'combined' dependence $f(r)$ in Fig. 2a and $\beta \approx 1.45$ for the first 15 points of the corresponding $p(f)$ dependence in Fig. 2b, in ~ 44% disagreement with formula (5). Somewhat unexpectedly, the correlation between the exponents α and β improves greatly and appears to be closest to that predicted theoretically for the cases $l = 11$–15 characterized with rather restricted statistical data (see also Fig. 2a and Fig. 2b).

Since the empirical data for α are more reliable, in Fig. 3 we show additionally the 'theoretical' $\beta(l)$ dependence calculated basing on the formula (5) and the empirical α data. Because of restricted accuracy of the data, our further analysis will deal with general qualitative tendencies in the $\alpha(l)$ and $\beta(l)$ evolution occurring with increasing word length, rather than the exact values of those exponents. In spite of all these limitations, our following conclusions drawn from the data displayed in Fig. 3 seem to be both grounded and important.

First, we observe a clearly visible decreasing trend for the rank–size exponent α with increasing word length l and an undoubted increase, under the same conditions, in the exponent β governing the frequency–probability dependence. Following from this tendency, considering a natural continuity of the $\alpha\,(l)$ and $\beta\,(l)$ functions and taking the formula (5) into account, one can reasonably suppose that the data corresponding to the smallest lengths l are to be described by the set of exponents $\alpha \to \infty$ (implying $f(r) \propto \exp(-ar)$) and $\beta \to 1$, rather than by the alternative set $\alpha \to 0$ (i.e., $f(r) \propto \log(r^{-1})$) and $\beta \to \infty$ (i.e., $p(f) \propto \exp(-bf)$, with b being a constant). This reasoning can help in resolving the problem of the $f(r)$ function for $l = 2$ in favour of formula (5), although a direct empirical confirmation would be desired.

As seen from Fig. 3, the Zipf's plots for the more common word lengths $l = 5$–10 reveal the values $\alpha \approx 1$ and $\beta \approx 2$. These are the standard exponents obtained in the most of computational linguistic studies of texts. On the other hand, the α parameter for the cases of $l = 12$–15 drops essentially and becomes less than one, while β increases above the value of two, contrary to common beliefs. As a result, one can pick up a slight hint at a 'crossover' to the logarithmic $f(r)$ function (see formula (3)) and an exponential lexical spectrum, $p(f) \propto \exp(-bf)$. Anyway, these cases also need more detailed empirical investigations. Notice also that a specific lexicon characterized with $l = 12$– 15 comprises a large portion of compound words and unique authorial proper names.

Finally, it is instructive to compare our results with those peculiar for the simplest 'Miller's monkey' random texts where all of the characters except for, perhaps, a blank as a word separator have the same probabilities (see, e.g., [8, 12]). For such a random text, the probabilities of any words of a given length l turn out to be the same. Then the analogues of the dependences shown in Fig. 2a and Fig. 2b should be uniform distributions, if only finite-size effects are neglected. This pattern is far from what we have really observed for the natural text. Only 'combined' $f(r)$ dependence for the whole set of the words of different lengths in the random text would acquire a shape of a power function, though distorted by multiple stairs present in the overall range of ranks (see, e.g., Ref. [13]).

Conclusions. Summing up, in the present work we have studied empirically and analyzed phenomenologically the rank–frequency dependences and the lexical frequency spectra typical for the groups of words in a natural text that have different lengths. Somewhat similar to the case of dividing a text into subsystems of N-grams with different lengths, we have proved that the first and the second Zipf's laws are better satisfied for the entire ensemble of our linguistic subsystems ($l = 1 \div l_{\max}$) rather than for the separate subsystems of the text (i.e., individual subsets of words with $l = 1$, $l = 2$, ...).

We have shown that, among different lexical subsystems, the subsystems referred to the middle-length words reveal the rank dependences and the probability density functions which are the closest to those typical for all the words in the text. In particular, their power-law exponents are close to the canonical values $\alpha \approx 1$ and $\beta \approx 2$. On the contrary, the subsystems with the smallest and largest word lengths are characterized by the exponents that contradict the known results for the overall word system. In particular, we hypothesize that the $f(r)$ and $p(f)$ functions at $l = 2$ can turn out to by close to exponential and power (with $\beta \sim 1$), respectively. At the same time, there are some reasons to believe that the rank–frequency and frequency–probability dependences for the largest word lengths l can be described respectively by the logarithmic and exponential functions.

To our opinion, it would be promising to further investigate the $f(r)$ and $p(f)$ dependences for the words with different lengths for much longer texts or corpora. Moreover, it would be desirable to compare comprehensively the dependences *f*(*r*) and *p*(*f*) for the natural texts with those obtained for the random texts, in which the letter probabilities are borrowed from the natural texts. Such investigations would help to understand which reasons and phenomena – purely statistical or fundamentally linguistic ones – underlie the fact that the Zipf's laws are better fulfilled for the whole systems of linguistic subsystems in the texts (cf. with the conclusions drawn in the work [3]).

One of the authors (O.S.K.) wishes to thank Assoc. Prof. Shuwar R. Ya. for stimulating our work in the field of computational linguistics and for his encouraging discussions.

References

1. *Egghe W.* On the law of Zipf-Mandelbrot for multi-word phrases // J. Amer. Soc. Inform. Sci. – 1999. – Vol. 50. – P. 233–241.
2. *Ha L. Q.*, *Sicilia-Garcia E. I.*, *Ming J.*, *Smith F. J.* Extension of Zipf's law to words and phrases // Proc. 19th International Conference on Computational Linguistics. – 2002. – Vol. 1. – P. 315–320.
3. *Ha L. Q.*, *Hanna P.*, *Ming J.*, *Smith F. J.* Extending Zipf's law to n-grams for large corpora // Artificial Intelligence Rev. – 2009. – Vol. 32. – P. 101–113.
4. *Ferrer i Cancho R.*, *Solé R. V.* Zipf's law and random texts // Adv. Complex Syst. – 2002. – Vol. 5. – P. 1–6.
5. *Newman M. E. J.* Power laws, Pareto distributions and Zipf's law // Contemporary Phys. – 2005. – Vol. 46. – P. 323–351.
6. *Гусейн-Заде С. М.* О распределении букв русского языка по частоте встречаемости // Проблемы передачи информации. – 1988. – Т. 24, №4. – С. 102–107.
7. *Kanter I.*, *Kessler D. A.* Markov processes: linguistics and Zipf's law // Phys. Rev. Lett. – 1995. – Vol. 74. – P. 4559–4562.
8. *Lü L.*, *Zhang Z.-K.*, *Zhou T.* Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes // Sci. Rep. – 2013. – Vol. 3. – 1082 (7 p.).
9. *Goldstein M. L., Morris S. A., Yen G. G.* Problems with fitting to the power-law distribution // Eur. Phys. J. B. – 2004. – Vol. 41. – P. 255–258.
10. *Corral A., Deluca A.* Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions // Acta Geophys. – 2013. – Vol. 61. – P. 1351–1394.
11. *Ferrer i Cancho R.*, *Solé R. V.* Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited // J. Quant. Linguist. – 2001. – Vol. 8. – P. 165–173.
12. *Li W.* Random texts exhibit Zipf's-law-like word frequency distribution // IEEE Trans. Inform. Theory. – 1992. – Vol. 38. – P. 1842–1845.
13. *Ferrer-i-Cancho R.*, *Elvevåg B.* Random texts do not exhibit the real Zipf's law-like rank distribution // PLoS One. – 2010. – Vol. 5. – e9411 (10 p.).

Стаття: надійшла до редакції 12.06.2015,
доопрацьована 18.06.2015,
прийнята до друку 23.06.2015.

РАНГОВІ ЗАЛЕЖНОСТІ ТА ЛЕКСИЧНІ ЧАСТОТНІ СПЕКТРИ ДЛЯ ПІДГРУП СЛІВ ТЕКСТУ З РІЗНИМИ ДОВЖИНАМИ

О. С. Кушнір, Л. Б. Іваніцький, М. Я. Максисько, С. В. Рихлюк

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

У роботі знайдено статистичний розподіл слів із англомовного тексту роману Дж. Толкіна “Володар Перснів” за їхніми довжинами. Вивчено “парціальні” залежності *f*(*r*) частоти від рангу і густини ймовірності частоти *p*(*f*) для підгруп слів окремих довжин (l = 1–16 букв), а також відповідні “об’єднані” залежності для слів усіх довжин. Установлено, що “парціальні” залежності загалом гірше описуються степеневими законами, відомими як закони Ціпфа, аніж “об’єднані”. Найближчими до степеневої є функції *f*(*r*) і *p*(*f*) для проміжних довжин слів l = 5–10, які описуються показниками $\alpha \approx 1$ для *f*(*r*) і $\beta \approx 2$ для *p*(*f*). Висловлено гіпотезу, що рангові залежності для найменших довжин можуть виявитися близькими до експоненційних $(\alpha \to \infty)$, схоже до функції *f*(*r*) для деяких східних мов із обмеженим словником, а рангова залежність і густина ймовірності для найбільших l – прямувати в границі відповідно до логарифмічної $(\alpha \to 0)$ і експоненційної $(\beta \to \infty)$ функцій.

Ключові слова: статистичні розподіли, степеневі розподіли, рангові залежності, комп’ютерна лінгвістика, природні мови, закон Ціпфа, розподіл слів за довжинами.

РАНГОВЫЕ ЗАВИСИМОСТИ И ЛЕКСИЧЕСКИЕ ЧАСТОТНЫЕ СПЕКТРЫ ДЛЯ ПОДГРУПП СЛОВ ТЕКСТА С РАЗЛИЧНЫМИ ДЛИНАМИ

О. С. Кушнир, Л. Б. Иваницкий, М. Я. Максысько, С. В. Рыхлюк

Львовский национальный университет имени Ивана Франко,
ул. Ген. Тарнавского, 107, 79017 Львов, Украина
o_kushnir@franko. lviv.ua

В работе найдено статистическое распределение слов из англоязычного текста романа Дж. Толкина “Властелин Колец” за их длинами. Изучены “парциальные” зависимости *f*(*r*) частоты от ранга и плотности вероятности частоты *p*(*f*) для подгрупп слов отдельных длин (l = 1–16 букв), а также соответствующие “объединенные” зависимости для слов всех длин. Установлено, что в общем случае “парциальные” зависимости хуже описываются степенными законами, известными как законы Ципфа, чем “объединенные”. Ближайшими к степенной являются функции *f*(*r*) и *p*(*f*) для промежуточных длин слов l = 5–10, описываемые показателями $\alpha \approx 1$ для *f*(*r*) и $\beta \approx 2$ для *p*(*f*). Выдвинута гипотеза, что ранговые зависимости для наименьших длин могут оказаться близкими к экспоненциальным $(\alpha \to \infty)$, похоже к функции *f*(*r*) для некоторых восточных языков с ограниченным словарем, а ранговая зависимость и плотность вероятности для наибольших l – стремиться в границе соответственно к логарифмической $(\alpha \to 0)$ и экспоненциальной $(\beta \to \infty)$ функциям.

Ключевые слова: статистические распределения, степенные распределения, ранговые зависимости, компьютерная лингвистика, естественные языки, закон Ципфа, распределение слов за длинами.