https://doi.org/10.15407/ujpe65.2.143

A.N. VASILEV, I.V. VASILEVA

Taras Shevchenko National University of Kyiv
(60, Volodymyrs'ka Str., Kyiv 01601, Ukraine; e-mail: vasilev@univ.kiev.ua)

PHYSICS BEYOND PHYSICS: APPLICATION OF PHYSICAL APPROACHES IN QUANTITATIVE LINGUISTICS

The application of physical methods to solve non-physical problems has been considered. In particular, the prospects of physical approaches in quantitative linguistics are analyzed. The difference between the physical and non-physical methods is illustrated by an example of already existing "classical" models. A few mathematical models which make it possible to determine the rank-frequency dependence for words in a frequency dictionary, as well as the dependence of the dictionary volume on the text length, are proposed. It is shown that the physical approaches and principles that are used in physics can also be successfully applied to create mathematical models in linguistics.

K e y w o r d s: physical theory, model, econophysics, sociophysics, quantitative linguistics.

Thy heart by one sole impulse is possess'd;
Unconscious of the other still remain!

Johann Wolfgang von GOETHE,
"Faust"

1. Introduction

The scope of problems tackled by physicists expands permanently, and physical methods become more and more involved in solving the problems that are not directly related to physics. This is not about isolated cases. The matter concerns a systematic approach in which economic, social, political, linguistic, and some other problems are formulated in the forms of models, similarly to what is a common practice in physics. For instance, such research directions as econophysics [1–8] and sociophysics [8–13] are currently quite familiar and acceptable to the physical community.

Moreover, relevant studies gain recognition among economists, sociologists, and political scientists – in other words, non-physical experts. This circumstance is not trivial, because the methodology of researches that is inherent, say, to econophysics is fundamentally different from the methods and models that are common to professional economists. At the same time, it is too early to talk about the total recognition of physical methods in non-physical scientific domains. The examples of researches given above remain within a scope of activity, where mainly physicists play the roles of both the driving force of researches and the "consumers" of the results obtained. This situation seems to be not quite right, especially if we take into account that the matter concerns approaches that are at least not worse than the modeling method considered as a standard in social sciences and humanities. As a confirmation of this statement, a number of papers devoted to the modeling and research of complex systems can be mentioned [14–18]. Their specificity consists in that it is sometimes rather difficult to determine which domain of knowledge the relevant system should be classified to. But despite that, the application of physical approaches brings about excellent results.

Besides econophysics and sociophysics, one of the research domains in which physical approaches and models can be successfully applied is quantitative linguistics [19–24]. Here, there is a considerable body of results obtained till now. Nevertheless, the search for effective ways to create new models remains a challenging task. In this work, we analyze the advantages of the physical modeling methods and outline directions in quantitative linguistics, where those methods can be applied.

2. Physical Approach to Modeling

Hence, what can be a motivation for the physical approaches to be applied when solving non-physical

problems, in particular, in quantitative linguistics? In order to answer this question, several important circumstances have to be taken into consideration. For instance, there are many interesting problems in which the application of a mathematical apparatus is required or allowed. This is no wonder, and the corresponding mathematical direction in linguistics has a long productive history. Accordingly, mathematical models, both proposed earlier and permanently emerging now, are successfully implemented in quantitative linguistics to describe a variety of systems and processes. At the same time, essential is not only the availability of that or another model, but also how it was created. Here, we are faced with what could be called a specific feature of the physical approach to the modeling.

As a rule, the models developed to describe physical phenomena or processes are based on a definite theory. This means that, at first, some rules or laws are formulated, in that or another way, for the interaction among the elements included in the examined system. The model is created only afterward, as a consequence of those laws. Even if the temporal sequence of the events is opposite (i.e. firstly either a model is created or a regression dependence is built on the basis of empirical data), ultimately there appears a theory that explains the corresponding mathematical relations and later becomes a basis for their derivation. As an example, we can point to the third Kepler law or the Stefan-Boltzmann law, which were first revealed experimentally and only afterward explained theoretically.

Unlike this physical approach, when the mathematical modeling is applied to the solution of linguistic and other problems, the relevant mathematical model is simply postulated or constructed proceeding from the convenience and the general appearance of the available "experimental" dependence. Models of this type do not always describe the corresponding system or process efficiently and completely. Why so? There are two important points to highlight. First of all, the absence of a theory in the basis of the model does not allow conclusions to be made about the essence of the mechanisms that are responsible for the resulting dependence. In effect, the model is descriptive in this case, which substantially reduces its value. There immediately arises a question as to whether the model is applicable, which in turn may call into question the reliability of the results obtained in its framework. Second, the specificity of a verification of the quantitative linguistics models on the basis of available data has to be taken into account. The matter is that, as a rule, the results of direct "measurements" are grouped and processed before their usage for the modeling [21–24]. In the absence of a basic theory, it is difficult (and sometimes impossible) to determine how the data grouping affects the character of the ultimate mathematical dependence. In other words, the model may turn out so "non-universal" that a change in the form of data presentation may qualitatively impact the behavior of the relevant functional dependence. This is a serious problem, and the way to solve it passes through the application of reasonable and universal principles determining the means for the creation of mathematical models. Those means are nothing else but the approaches developed and used by physicists in order to successfully model systems of various origins, including the linguistic ones (see, e.g., works [25–30] and references therein).

3. Models of Quantitative Linguistics

Before proceeding to the analysis of the ways used for the implementation of physical approaches to the solution of linguistic problems, let us consider some already existing "classical" models that are used in quantitative linguistics. Historically, Zipf's law was among the first ones that appeared in linguistics [31–34]. It relates the frequency f at which a certain word appears in the text to the rank n of this word. More specifically, the number of different words in a large volume of the text is determined, and the number of occurrences in the text is calculated for each word. Conventionally, this parameter is called the frequency of the word (or the word occurrence) in the text. The words are ranged in the descending order of their occurrence in the text. The rank is an ordinal number of the word in that sequence (i.e., a word with rank 1 has the largest number of occurrence in the text). According to Zipf's law, this dependence has to be power-law,

$$f(n) = \frac{A}{n^{\alpha}}. \tag{1}$$

Here, A is a certain non-universal constant, and α is the power exponent (index). The calculation of the latter usually comprises the main purpose of the research, because there is evidence that the value of

this index is close to unity for many languages and various non-specialized texts [21, 35]. The causes for and consequences of α deviations from unity are the subject of a number of separate studies. Relation (1) is "empirical", because it was found by processing a large amount of linguistic data. Moreover, it is only obeyed within a certain interval of word rank values. It is evident that, according to the definitions of the quantities f and n, the following relation has to be satisfied:

$$V = \sum_{n=1}^{N} f(n), \tag{2}$$

where N denotes the number of different words (lexemes) in the text, and V stands for the text volume, i.e., the total number of words in the text.

It should also be noted at once that the validity of relation (1) means a linear relation between the quantities $\ln f$ and $\ln n$:

$$\ln f = B - \alpha \ln n, \tag{3}$$

where $B = \ln A$. As an illustration, Fig. 1 demonstrates the rank versus the frequency relationship for the text of *Notes of Ukrainian Crazy* by Lina Kostenko [36]. This text contains 85227 words, and the number of lexemes (different words) equals 14796. On the basis of the data presented at the information resource *www.mova.info*, we managed to calculate the values $\alpha \approx 0.948$ and $B \approx 8.796$ of the parameters in dependence (3) used for the approximation. The coefficient of determination is $R^2 \approx 0.964$. In the considered case, the approximation was performed within the entire interval of word rank values. If the analysis is confined only to the interval, where the dependence is substantially linear (in particular, at $4 \leq \ln n \leq 7$), then the value $\alpha \approx 0.995$ is obtained, with the coefficient of determination $R^2 \approx 0.997$.

It is easy to see that, as the rank increases, the number of words with the same frequency also increases. Using the notation $m(f)$ for the number of words that occur with the frequency f, the "spectral" analog of the rank distribution is the law [21]

$$m(f) = \frac{M}{f^{1+\gamma}}, \tag{4}$$

where γ and M are the distribution parameters. In particular, for the text indicated above, we have

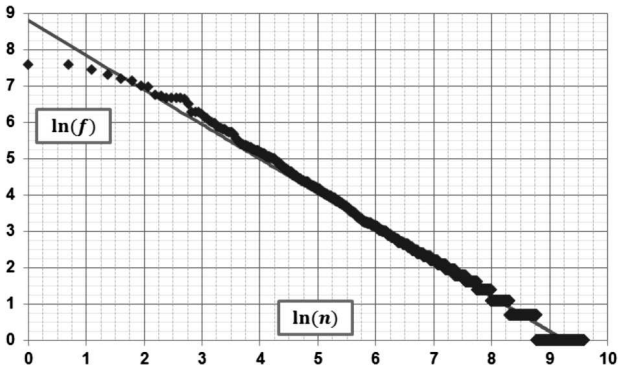

Fig. 1. Rank versus the frequency for the text of *Notes of Ukrainian Crazy* by Lina Kostenko [36]. Symbols correspond to "experimental" values. Solid curve is an approximation based on Zipf's law. The power exponent in the distribution $\alpha \approx 0.948$

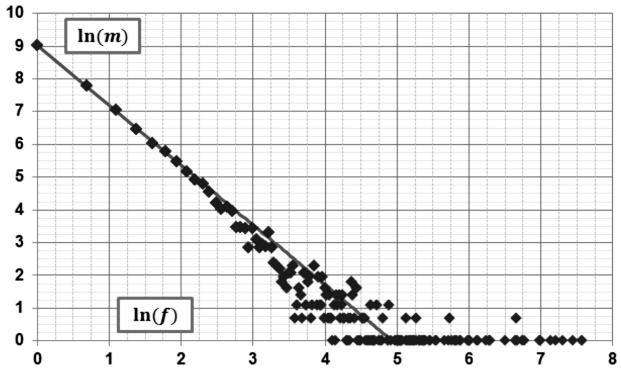

Fig. 2. "Spectral" word distribution for text [36]. Symbols correspond to "experimental" values. Solid curve is an approximation based on law (4). The power exponent in the distribution $\gamma \approx 0.825$

$\gamma \approx 0.825$, $M \approx 9.015$, and the coefficient of determination $R^2 \approx 0.889$. The corresponding dependence $\ln m$ versus $\ln f$ is plotted in Fig. 2.

Note that relations (1) and (4) are sometimes referred to as the first and second Zipf's laws, respectively, and there exists a non-trivial connection between the corresponding relations (see, e.g., work [37]).

Another example of a dependence that is rather often used in practice is the dependence of the dictionary size N (number of different words in the dictionary text) on the text volume V (the number of all words in the text). This dependence is nonlinear at large V, and there is no general universal formula in this case. There are various approaches to this problem, in which the form of the approximation function is chosen *a priori* [30, 38–49].

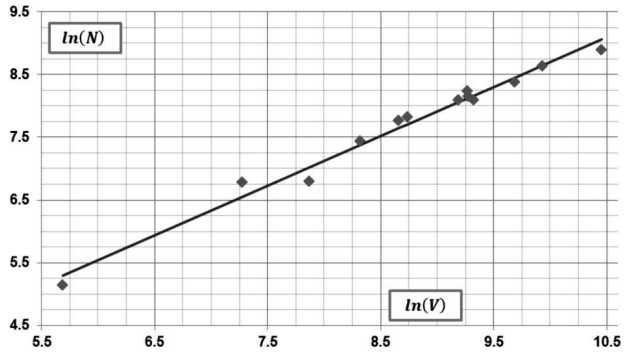

Fig. 3. Dependence of the dictionary size on the text volume for works of Taras Prokhas'ko. Symbols correspond to actual values, and the solid curve is an approximation on the basis of dependence (5). The parameter $\beta \approx 0.788$

On the basis of general considerations, an attempt can be made to describe the corresponding functional dependence by the power law

$$N(V) = kV^{\beta} \tag{5}$$

or, in terms of logarithms,

$$\ln N = \beta \ln V + K, \tag{6}$$

where β, k, and $K = \ln k$ are model parameters. An example of the application of this dependence to describe real data is exhibited in Fig. 3. The figure illustrates the relationship between the logarithmized dictionary size, $\ln N$, and the logarithmized text volume, $\ln V$, for a number of works by Taras Prokhas'ko (the data were obtained from the information resource *www.mova.info*). The calculations gave the values $\beta \approx 0.788$ and $K \approx 0.816$ for the distribution parameters and $R^2 \approx 0.982$ for the coefficient of determination.

All the relationships and the results of "empirical" data processing presented above throw light on two problems: a technical problem and a methodological one. The former arises because each of relations (1), (4), and (5) describes data only in a certain interval. For example, it is well known that Zipf's law (1) is inapplicable to the distribution of words with low and high ranks. There are some difficulties with the application of "spectral" law (4) to the distribution of high-frequency words [21]. The applicability of dependence (5) is also restricted, because, e.g., the evident equality $N = 1$ must be obeyed at $V = 1$.

All this means that the functional dependences (1), (4), and (5) are approximate and should be specified in principle, but how? Here, we are faced with a methodological problem. In general, the most complicated expression for the approximation dependence is chosen in this case, and the parameters of this expression are determined on the basis of "empirical" data. However, no clear criteria have been currently developed to determine which approximation should be selected. So, it is here where the methods used by physicists can be useful.

4. Implementation of Physical Approaches in Linguistics

The idea is extremely simple. It is based on the fact that the approximation dependence can be obtained by solving a differential equation. The equation, in turn, is written on the basis of general concepts about the character of the processes that "govern" the analyzed dependence. For example, formula (1) for Zipf's law can be obtained as a solution of the first-order differential equation

$$\frac{df}{f} = -\alpha \frac{dn}{n}. \tag{7}$$

According to it, the relative change of the word frequency is proportional to the relative change of the word rank. It is easy to guess that a similar equation can serve as a basis for obtaining laws (4) and (5) with an accuracy of notations used in the corresponding relationship. In effect, this means that, analogously to the proportionality between the relative change of the word occurrence frequency and the relative change of the word rank, the relative change in the number of words occurring with a certain frequency is proportional to the relative frequency change, and the relative change in the number of different words in the text is proportional to the relative change of the text volume.

Those laws can be summarized and made a bit more universal. As an initial point, let us assume that two parameters (like the number of different words in the text and the text volume) change in such a way that, by nonlinearly transforming them separately, the corresponding changes can be made proportional to each other. Using the notations x and y for those parameters, our assumption can be written down in

the form of a differential equation

$$\phi(y)\,dy = \psi(x)\,dx, \tag{8}$$

where $\phi(y)$ and $\psi(x)$ are some functions, which are unknown *a priori*. We can estimate them by expanding them into Taylor series in negative powers of the corresponding argument (in order to obtain the known laws already in the first approximation). In particular, if the linear (in the reciprocal argument) approximation is used for the function $\phi(y)$ and the cubic approximation for the function $\psi(x)$, then, taking into account that any of the coefficients can be chosen arbitrarily, we obtain the following equation:

$$\frac{dy}{y} = \left(\frac{a}{x} + \frac{b}{x^2} + \frac{2c}{x^3}\right) dx, \tag{9}$$

where the parameters a, b, and c are determined by approximating the "empirical" data. The solution of this equation looks like

$$y(x) = y_0 x^a \exp\left(-\frac{b}{x} - \frac{c}{x^2}\right). \tag{10}$$

It serves as a basis for constructing an approximation dependence (with four parameters: a, b, c, and y_0). In terms of the new variables $z = \ln y$ and $t = \ln x$, the sought approximation dependence reads

$$z(t) = at - b\exp(-t) - c\exp(-2t) + d, \tag{11}$$

where $d = \ln y_0$.

Hence, we obtained a dependence with logarithmic variables. It can be regarded as containing the exponential corrections to the linear law. Moreover, the approximated dependences are rather monotonic, and several parameters are used for the approximation. Therefore, when calculating those parameters, not only the least squares method (or another criterion) can be applied, but some additional constraints can also be imposed, which is important for the solution of linguistic problems.

As an example, Fig. 4 demonstrates the approximation results for the word rank distribution in Lina Kostenko's text [36], but now the analysis is based on expression (11). In particular, this is a relation between the logarithms of the word frequency, $\ln f$, and the word rank, $\ln n$. The corresponding distribution parameters are as follows: $a \approx -0.951$, $b \approx 1.531$, $c \approx -0.285$, and $d \approx 8.822$. The coefficient of determination $R^2 \approx 0.964$. At calculations, two additional conditions were monitored: (i) the value of the approximation function at the starting point had to coincide with the corresponding "experimental" value and (ii) the derivative could not exceed zero.

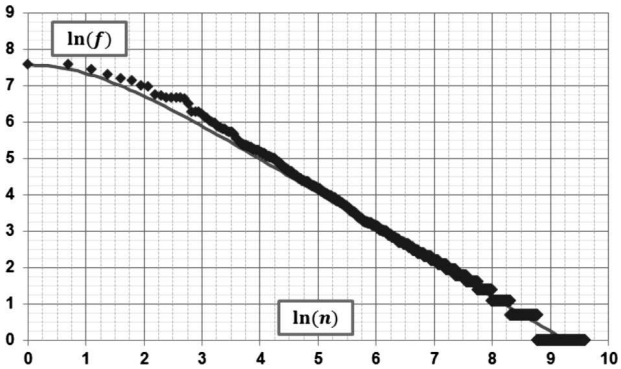

Fig. 4. Rank versus frequency for text [36]. Symbols correspond to "experimental" values. Solid curve is an approximation based on law (11)

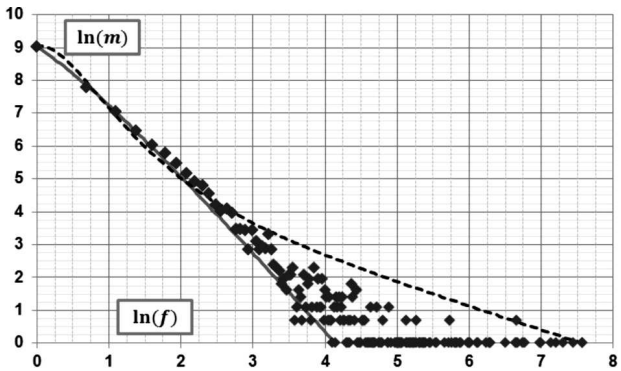

Fig. 5. Spectral word distribution for text [36]. Symbols correspond to "experimental" values. The solid curve corresponds to the approximation over the minimum values, and the dashed curve to the approximation over the maximum values, the both on the basis of law (11)

When modeling the spectral distribution, we are faced with a problem that the corresponding function becomes essentially ambiguous, as the word frequency increases. In this case, for example, we can do the approximation after fixing the values of the first and last points in the dependence. In Fig. 5, the results of the corresponding approximation for the text [36] are shown. If the approximation is performed after fixing the values of the first point (the zero value for the logarithm of the word frequency $\ln f$) and the

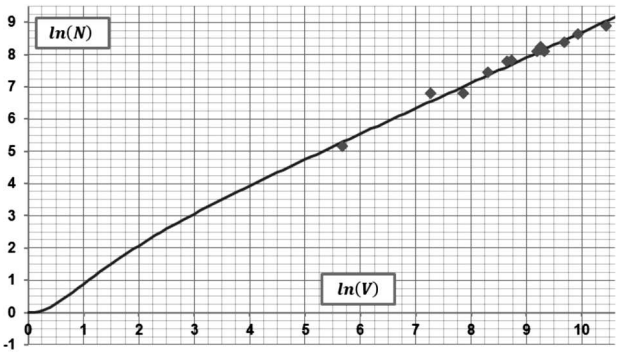

Fig. 6. Dependence of the dictionary size on the text volume for works of Taras Prokhas'ko. Symbols correspond to "experimental" values. Solid curve is an approximation based on law (11)

first point at which the logarithm of the number of words with a given frequency equals zero ($\ln m = 0$), then we obtain the following approximation parameter values: $a \approx -2.424$, $b \approx 1.081$, $c \approx -0.045$, and $d \approx 10.061$. The coefficient of determination $R^2 \approx 0.898$. But if we fix the last point for which the logarithm of the number of words with the corresponding frequency equals zero, the approximation gives the following parameter values: $a \approx -0.708$, $b \approx -8.643$, $c \approx 4.981$, and $d \approx 5.362$. The coefficient of determination $R^2 \approx 0.290$. In the latter case, one can hardly talk about the qualitative approximation, but rather about a curve that describes the limiting values for the corresponding dependence.

Finally, Fig. 6 illustrates the results of approximation of the relationship between the number of lexemes in the text (the dictionary size) and the text volume on the basis of expression (11) applied to Taras Prokhas'ko's texts (*www.mova. info*). At calculations, the following additional restrictions were used: (i) for texts consisting of one word, the corresponding dictionaries were considered to also consist of one word; and (ii) the number of words in the dictionary could not be negative. Provided the indicated additional conditions, the following values of the approximation parameters were obtained: $a \approx 0.786$, $b \approx 2.670$, $c \approx -1.835$, and $d \approx 0.835$. The coefficient of determination $R^2 \approx 0.982$. In this case, the effect of using a nonlinear dependence is insignificant, but this dependence made it possible to build another dependence that gave correct values even for texts with small volumes.

5. Conclusions

To summarize, an approach has been proposed which enables the creation of general approximate dependences that can be applied to the mathematical modeling in quantitative linguistics. The basic idea consists in the application of a certain differential equation to the description of a process or a relationship. The approximation dependence is based on the general solution of this differential equation. In addition to the immediate advantage that an approximate dependence can be obtained, this approach allows the classification of linguistic models and the determination of their scope of applicability to be made, which can be a crucial factor from the methodological standpoint. The examples presented in the article to illustrate this approach give grounds to assume that its application can be promising in other cases as well. Furthermore, the proposed methodology may provide additional confirmation and substantiation of the results obtained in the framework of alternative theories; as it occurs, e.g., with various explanations of Zipf's law [21, 35, 50].

The authors are sincerely grateful to Prof. Nataliya Darchuk and her colleagues for the development and support of the information resource www.mova.info, which was used in this work. The authors are also extremely thankful to the Referees for their advice and suggestions, which enabled us to improve the article.

1. D. Walker. Economics and social physics. *Econom. J.* **101**, 615 (1991).
2. R. Mantegna, H. Stanley. *An Introduction to Econophysics* (Cambridge Univ. Press, 2000).
3. J. McCauley. *Dynamics of Markets: Econophysics and Finance* (Cambridge Univ. Press, 2004).
4. F. Jovanovic, C. Schinckus. Econophysics: A new challenge for financial economics. *J. Hist. Econ. Thought* **35**, 319 (2012).
5. Y. Gingras, C. Schinckus. The institutionalization of econophysics in the shadow of physics. *J. Hist. Econ. Thought* **34**, 109 (2012).
6. C. Schinckus, F. Jovanovic. Towards a transdisciplinary econophysics. *J. Econ. Method.* **20**, 164 (2013).
7. D. Sornette. Physics and financial economics (1776–2014): Puzzles, Ising and agent-based models. *Rep. Progr. Phys.* **77**, 1 (2014).
8. B. Chakrabarti, A. Chakraborti, A. Chatterjee. *Econophysics and Sociophysics: Trends and Perspectives* (Wiley-VCH, 2006).

9. S. Galam, Y. Gefen, Y. Shapir. Sociophysics: A mean behavior model for the process of strike. *J. Math. Sociol.* **9**, 1 (1982).
10. S. Galam. *Sociophysics: A Physicist's Modeling of Psycho-Polytical Phenomena* (Springer, 2012).
11. D. Stauffer. A biased review of sociophysics arXiv: 1207.6178v1.
12. C. Castellano, S. Fortunato, V. Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591 (2009).
13. S. Galam. Sociophysics: A review of Galam models. arXiv: 0803.1800.
14. B. Berche, C. von Ferber, T. Holovatch, Yu. Holovatch. Transportation network stability: a case study of city transit. *Adv. Complex Syst.* **15**, 1, 1250063 (2012).
15. Y. Holovatch, V. Palchykov. Complex networks of words in fables. In: *Maths Meets Myths: Complexity-Science Approaches to Folktales, Myths, Sagas, and Histories.* Edited by R. Kenna, M. Mac Carron, P. Mac Carron (Springer, 2016).
16. Yu. Holovatch, R. Kenna, S. Thurner. Complex systems: physics beyond physics. *Eur. J. Phys.* **38**, 023002 (2017).
17. Yu. Golovach, M. Dudka, V. Blavatska, V. Palchykov, M. Krasnytska, O. Mryglod. *Statistical Physics of Complex Systems.* Preprint ICMP-17-06U (Lviv, 2017) (in Ukrainian).
18. Yu. Golovach, M. Dudka, V. Blavatska, V. Palchykov, M. Krasnytska, O. Mryglod. Statistical physics of complex systems in the world and Lviv. *Zh. Fiz. Dosl.* **22**, 2801 (2018) (in Ukrainian).
19. G. Altmann, R. Köhler. "Language Forces" and synergetic modelling of language phenomena. *Glottometrika* **15**, 62 (1996).
20. R. Köhler. Synergetic linguistics. In *Quantitative Linguistics. An International Handbook* (Walter de Gruyter, 2005), p. 760.
21. Yu. Tuldava. *Problems and Methods of Quantitative-Systemic Research of Lexicon* (Valgus, 1987) (in Russian).
22. R. Piotrovskii, K. Bektaev, A. Piotrovskaya. *Mathematical Linguistics* (Vysshaya Shkola, 1977) (in Russian).
23. R. Piotrovskii. *Linguistic Synergetics: Basic Assumptions, First Results, Prospects* (St. Petersburg State Univ., 2006) (in Russian).
24. V.V. Levitskii. *Quantitative Methods in Linguistics* (Ruta, 2005) (in Russian).
25. Yu. Golovach, V. Palchykov. Mykyta fox and the language network. *Zh. Fiz. Dosl.* **11**, N 1, 22 (2007) (in Ukrainian).
26. A.A. Rovenchak, S. Buk. Defining thermodynamic parameters for texts from word rank-frequency distributions. *J. Phys. Stud.* **15**, 1005 (2011).
27. A.A. Rovenchak, S. Buk. Application of a quantum ensemble model to linguistic analysis. *Physica A* **390**, 1326 (2011).
28. A. Rovenchak, S. Buk. Part-of-speech sequences in literary text: Evidence from Ukrainian. *J. Quant. Linguist.* **25** (1), 1 (2018).
29. O.M. Vasilev, O.V. Chalyi, I.V. Vasileva. On "exotic" problems in physics, Winnie the Pooh, and Zipf's law. *Zh. Fiz. Dosl.* **17**, 1001 (2013) (in Ukrainian).
30. A. Vasilev, I. Vasileva. Text length and vocabulary size: Case of the Ukrainian writer Ivan Franko. *Glottometrics* **43**, 1 (2018).
31. G. Zipf. *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949).
32. G. Zipf. *The Psycho-Biology of Language* (Addison-Wesley, 1935).
33. W. Li. Zipf's law everywhere. *Glottometrics* **5**, 14 (2002).
34. I.-I. Popescu, G. Altmann, R. Köhler. Zipf's law another view. *Qual. Quant.* **44**, 713 (2010).
35. V. Palchykov. *Scale-Free and Small-World Effects in Complex Networks.* Ph.D. thesis (Lviv, 2010) (in Ukrainian).
36. L. Kostenko. *Notes of Ukrainian Crazy* (A-ba-ba-ga-la-maga, Lviv, 2014) (in Ukrainian).
37. I. Moreno-Sánchez, F. Font-Clos, Á. Corral. Large-scale analysis of Zipf's law in English texts. *PLoS ONE* **11,** e0147073 (2016).
38. G. Thomson, J.R. Thompson. Outline of a measure for the quantitative analysis of writing vocabularies. *Brit. J. Psychol.* **8**, 52 (1915).
39. G. Herdan. *Type-Token Mathematics: A Textbook of Mathematical Linguistics* (Gravenhage, 1960).
40. J. Tuldava. The statistical structure of a text and its readability. In *Quantitative text analysis* (Wissenschaftlicher Verlag Trier, 1993).
41. J. Tuldava. On the relation between text length and vocabulary size. In: *Methods in Quantitative Linguistics* (Wissenschaftlicher Verlag Trier, 1995), p. 131
42. E. Panas. The generalized torquist: Specification and estimation of a new vocabulary-text size function. *J. Quant. Linguist.* **8**, 233 (2001).
43. E. Panas, A.N. Yannacopoulos. Stochastic models for the lexical richness of a text: Qualitative results. *J. Quant. Linguist.* **11**, 251 (2004).
44. G. Wimmer. The type-token relation. In *Quantitative Linguistics. An International Handbook* (Walter de Gruyter, 2005), p. 361.
45. R. Köhler. Synergetic linguistics. In *Quantitative Linguistics. An International Handbook* (Walter de Gruyter, 2005), p. 760.
46. F. Fan. Text length, vocabulary size and text coverage constancy. *J. Quant. Linguist.* **20**, 288 (2013).
47. M. Kubát, J. Milička. Vocabulary richness measure in genres. *J. Quant. Linguist.* **20**, 339 (2013).
48. D. Mitchell. Type-token models: a comparative study. *J. Quant. Linguist.* **22**, 1 (2015).
49. F. Fan, Y. Yang, W. Yaqin. The probability distribution of textual vocabulary in the English language. *J. Quant. Linguist.* **23**, 49 (2016).
50. S. Thurner R. Hanel, B. Liu, B. Corominas-Murtra. Understanding Zipf's law of word frequencies through sample-space collapse in sentence formation. *J. R. Soc. Interface* **12**, 20150330 (2015).

Received 15.10.19.
Translated from Ukrainian by O.I. Voitenko

О.М. Васильєв, І.В. Васильєва

ФІЗИКА ЗА МЕЖАМИ
ФІЗИКИ: ФІЗИЧНІ ПІДХОДИ
В КВАНТИТАТИВНІЙ ЛІНГВІСТИЦІ

Р е з ю м е

В статті розглядається проблема використання фізичних методів для розв'язання задач нефізичного характеру. Зокрема, аналізуються перспективи застосування фізичних підходів в кількісній (квантитативній) лінгвістиці. Різниця між фізичними та нефізичними способами моделювання ілюструється на прикладі уже існуючих "класичних" моделей. Також пропонуються математичні моделі, котрі дозволяють встановлювати рангово-частотну залежність для слів у частотному словнику та залежність розміру словника від об'єму тексту. Показано, що підходи і принципи, котрі є характерними для фізики, можуть бути з успіхом задіяні при створенні математичних моделей у лінгвістиці.