https://doi.org/10.15407/ujpe66.5.373

N.P. DARCHUK, I.V. VASILEVA, A.N. VASILEV

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

VECTOR MODEL FOR THE TEXT STYLE ANALYSIS

The application of physical approaches to the analysis of the authorial styles of Ukrainian writers has been considered. A model, where the literary styles are described in a multidimensional space with the help of unit vectors, is proposed. The numerical characteristic of the style is the scalar product of the corresponding vector and a vector that determines the general style for a group of authors. This parameter is shown to depend linearly on author's rank. This behavior confirms the hypothesis of joining the majority, according to which an author, when selecting his/her literary style, takes the style of his/her successful colleagues into account.

K e y w o r d s: physics of complex systems, quantitative linguistics, vector, energy, distribution.

The literature of every nation is the most perfect mirror of its life.

Ivan FRANKO

1. Introduction

Physics is permanently expanding the scope of its application. A striking example is a direction called the physics of complex systems (see, e.g., works [1–5] and the references therein). In the framework of this approach, systems of different nature are analyzed from the same position. The list of problems that have been successfully solved and are being solved using the methods of the physics of complex systems contains the analysis of transport systems [6–8], socioeconomic links [9–16], text structures and characteristics [17–27], and much more other things.

There are several important circumstances. First, as was mentioned above, the researched systems are completely different in their nature. Sometimes, it is rather difficult to determine which field of knowledge the relevant scope of problems belongs to. Second, such systems are studied making use of the methods that are inherent in physics (mainly, statistical physics), because those methods are characterized by the application of the appropriate mathematical apparatus. Furthermore, physical approaches also possess a powerful methodological arsenal which enables one to develop new models. The latter circumstance is extremely important, because it is the methodological aspects that come to the force, when the matter concerns interdisciplinary studies. The advantage of this approach consists in that original models allow not only the behavior and characteristics of the examined system to be analyzed, but also those parameters to be identified that should be determined on the basis of statistical data.

In this work, the physical methods are applied to analyze the authorial styles of Ukrainian writers. In *quantitative linguistics* [28–30], the problems of this kind have the highest priority (see, e.g., works [31–36] and the references therein). One of the ways to tackle them is based on the application of physical methods together with the corresponding categorical terminology.

2. Formalization of the Problem

First of all, let us formalize the problem. Let a certain number of words (the *basic set*) be used in a literary language. Any literary work can be considered as a sequence of words (with the corresponding grammatical changes) taken from the basic set. Every author is characterized by a set of texts. The latter can serve as a basis on which one can calculate the parameters that characterize the *style* inherent in this author. In particular, if the words from the basic set are arranged, e.g., alphabetically, then the works of the examined author could be described by a vector, the components of which determine the number of applications of every word in author's texts. However, the dimension of such a vector would be rather substantial, which disables

the usage of this approach in practice. More productive seems to be the approach where the basic set of words is significantly confined. For example, the set can only contain words that belong to a definite category or concept. Just this approach was used in this work.

Several points that are important for understanding the features of the proposed approach should be pointed out. First, it concerns the definition of authorial style. Generally speaking, the concept of style is quite broad and includes a set of attributes that characterize the manner of the writer and distinguish his/her works among the works of other authors [37]. Therefore, the authorial style undoubtedly cannot be described completely by means of a definite vector. Such a vector should only be considered as one of the characteristics of the authorial style. Second, as was also mentioned above, a limited group of words is used when constructing a vector that characterizes the authorial style. This group of words is a subset of the basic set. Therefore, the vector constructed on its basis can be interpreted as a projection of the vector constructed on the basis of the basic set onto the corresponding subset. In effect, this fact means that, by choosing different subsets of the basic set, we obtain different characteristics of the authorial style which are related to that or another particular topic in author's works.

Let us assume that the basic set consists of N words, and the number of authors equals M. Let $a_n^{(m)}$ denote how many times the n-th word ($n = 1, 2, ..., N$) was used by the m-th author ($m = 1, 2, ..., M$). Vectors $\mathbf{e}_m$ are defined to have the components

$$e_n^{(m)} = \frac{a_n^{(m)}}{\sqrt{\sum_{n=1}^{N} (a_n^{(m)})^2}}. \tag{1}$$

The scalar product of two arbitrary vectors $\mathbf{e}_m$ and $\mathbf{e}_k$ is determined by the formula

$$\mathbf{e}_m \cdot \mathbf{e}_k = \sum_{n=1}^{N} e_n^{(m)} e_n^{(k)}, \tag{2}$$

so that the introduced vectors have a unit length,

$$|\mathbf{e}_m|^2 \equiv \mathbf{e}_m \cdot \mathbf{e}_m = 1. \tag{3}$$

The vector $\mathbf{e}_m$ ($m = 1, 2, ..., M$) determined in such a way will be interpreted as one describing the literary style of the m-th author.

Let us also define the vector $\mathbf{a}$ with the components a_n ($n = 1, 2, ..., N$) that are calculated as follows:

$$a_n = \frac{A_n}{\sqrt{\sum_{n=1}^{N} A_n^2}}. \tag{4}$$

where the parameters A_m are defined by the formula

$$A_n = \sum_{m=1}^{M} a_n^{(m)}. \tag{5}$$

Hence, the vector $\mathbf{a}$ describes the general style, which is calculated on the basis of all the texts taken together.

It should be noted that the scalar product of two vectors (in our case, this is a value within the interval from 0 to 1) can be interpreted as a parameter characterizing the similarity of the authorial styles (an analog of the spin "interaction energy" in a model similar to the Ising model [38, 39]). In effect, the matter concerns the adaptation of a technique that is widely used in linguistic researches and is based on the calculation of the similarity coefficient (see, e.g., work [40]). In particular, if the scalar product equals unity, then the styles coincide with each other. If the product is zero, then the non-zero components of the first vector correspond to the zero components of the second vector and vice versa. This means that the authors whose styles are described by such vectors use completely different sets of words (from the basic set), with no "common" ones among them. In particular, the scalar product of the vectors $\mathbf{a}$ and $\mathbf{e}_m$ determines how much the style of the m-th author differs from the general style. The quantity

$$\varepsilon_m = \mathbf{a} \cdot \mathbf{e}_m \tag{6}$$

can be interpreted as a numerical characteristic that evaluates the style of the m-th author with respect to the general style. The distribution of the ε_m-values within a group of writers may seem to be important. However, in order to solve this issue, it is necessary to be able to describe the behavior of agents (writers), when they select their own styles. Here, an analogy can be traced with how the elements of a statistical system become distributed over the energy levels in an external field, with the role of the latter being played by the vector $\mathbf{a}$. Accordingly, the behavior of agents is analogous to additional restrictions imposed on the system (like the Pauli principle [41, 42]).

3. Effect of Joining the Majority

An important question is: "Are there any laws or principles that govern the distribution of the authorial style characteristics?" We proceed from the hypothesis that the society demonstrates demands for certain literary styles or topics, and the popularity of an author depends on how his/her style meets the expectations of the society. Therefore, it is reasonable to assume that some (not all) authors choose, consciously or subconsciously, their own styles on the basis of the styles inherent in other popular authors. On the other hand, the desire of every author to have his/her own unique style seems natural. If the authors are considered as social agents, then every of them must make a certain choice with respect to his/her own style. Different authors may follow different principles or approaches when choosing the style, and this fact does not allow us to talk about universal laws for the behavior of social agents. At the same time, the problem may concern the identification of the groups or communities of authors demonstrating common mechanisms to realize their own social choice.

There are several theories and approaches describing how agents in socio-economic systems make their choices. One of the behavioral algorithms is called the "effect of joining the majority" [43]. Its essence consists in that the agent behaves similarly to the majority of other agents. In the context of the researched model, such a behavior can be interpreted as an attempt of the m-th writer ($m = 1, 2, ..., M$) to maximize the value of the parameter ε_m. Actually, this means that the vectors $\mathbf{e}_m$ and $\mathbf{a}$ must be co-directed. However, the absolute coincidence of the styles will make the authors indistinguishable for the perception of their readers. Therefore, the difference between the authorial styles has to be appreciable. Just because the styles must be different, this effect cannot be reduced to the principle of dominant joining in the well-known Simon model [44]. Hence, we may expect to obtain new results.

Let $\Delta\varepsilon$ denote the minimal difference between the values of the parameters ε_m, when the readers still can catch the difference between the styles of different authors. With the help of the physical terminology, the situation can be interpreted as follows. There is a set of "energy" levels ε_m, and the difference between those levels equals $\Delta\varepsilon$. Different particles playing the role of agents (writers) must be located at different levels. In fact, we deal here with something like the Pauli principle in physics.

Without any loss of generality, we may assume that the authors are so enumerated that $\varepsilon_1 \geq \varepsilon_2 \geq \geq ... \geq \varepsilon_M$. Then, taking all the aforesaid into account, we have

$$\varepsilon_m = \varepsilon_0 - m\,\Delta\varepsilon \tag{7}$$

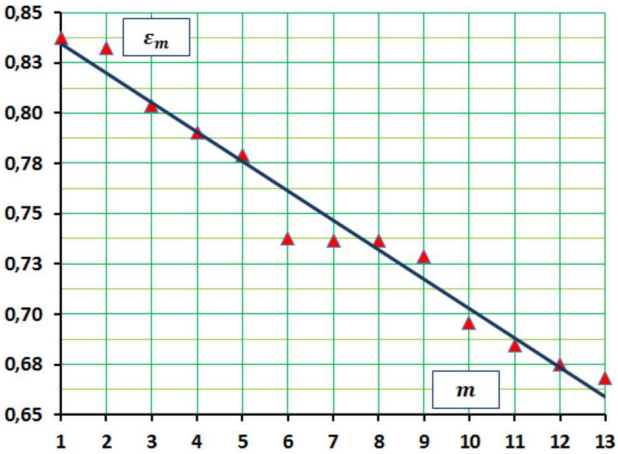
for all $m = 1, 2, ..., M$, where ε_0 is the model parameter. Equation (7) takes into account that the vector $\mathbf{e}_1$, which corresponds to the highest "energy" level ε_1, and the vector $\mathbf{a}$ can be not co-directed, because the latter does not describe a real style, being only a result of the averaging.

Therefore, if the "joining effect" does exist, and if we range the writers according to the values of the parameters ε_m, we must obtain a linear dependence of ε_m on author's rank m.

4. Empirical Results

Let us test the validity of the model proposed above on the basis of the available "experimental" results. In particular, let us analyze the information concerning the application of words belonging to the "human being/person" category by a group of Ukrainian poets/poetesses (*V. Stus, I. Drach, S. Yovenko, L. Kostenko, M. Vingranovs'kyi, M. Vorobyov, I. Zhylenko, O. Zabuzhko, V. Kolomiets', T. Kolomiets', B. Oliinyk, D. Pavlychko,* and *L. Skyrda*) with the parameter values $M = 13$ and $N = 115$. The analysis was performed on the basis of their poetic works written in the 1960s, 1970s, and 1990s. The volume of the text sample for each poet/poetess amounted to 20000 different words. Table contains the information on the number of lexemes from the "human being/person"category used by various authors, the number of lexico-semantic uses (LSU), and the total number of lexico-semantic forms in the texts (the dictionary) [45–47].

Our choice of the indicated group of poets and poetesses was made due to several reasons. In particular, it was necessary to identify authors belonging to a definite compact group, the behavior of which could be described in terms of the proposed model of joining the majority. The group members did not have to belong to a definite short time interval. On the contrary, if the process of joining the group is considered as the occupation of a certain vacancy in the "energy



Dependence of the parameter ε_m on author's rank m. The correspondence between the author and his/her rank is shown in Table

spectrum" and taking into account that the vacancies can be occupied in an arbitrary order, then the time required to fill the whole "spectrum" should last for a rather long period, which may include several generations of authors. Finally, the availability of the group and whether a particular author belongs to it are determined by a direct verification. In other words, our aim is not to prove the existence of a certain regularity in the distribution of the authorial style characteristics for any group of authors. Instead, the aim is to find a group of authors who satisfy a certain regularity. It was found empirically that the indicated poets and poetesses belong to the same group, which can be described in the framework of the model of joining the majority.

Data about the writers

Rank	Writer	LSU	Lexemes	Dictionary
1	V. Kolomiets'	341	31	4188
2	T. Kolomiets'	181	16	3536
3	L. Skyrda	252	19	3338
4	I. Drach	286	25	3818
5	B. Oliinyk	466	42	3463
6	M. Vingranovs'kyi	197	24	3252
7	S. Yovenko	272	25	3186
8	L. Kostenko	281	26	4040
9	M. Vorobyov	196	15	3193
10	D. Pavlychko	194	19	3830
11	O. Zabuzhko	128	13	4181
12	I. Zhylenko	317	29	3802
13	V. Stus	210	21	3664

The following two important circumstances should be pointed out. First, the choice of the "human being/person" group determines the analyzed subset for the style vector. From this viewpoint, it is quite acceptable to choose another group and analyze the corresponding subset. However, the new group should be rather general and common for lots of authors. Second, after the group for determining the subset has been selected, the authors are chosen according to the criterion of how the relevant topic is dealt with in their works. For instance, the number of applications of lexemes from the selected word group can be taken as a selection criterion.

On the basis of the available empirical data [45–47], the vectors describing authorial styles were constructed in the 115-dimensional space ($N = 115$) for all authors, and the vector $\mathbf{a}$ describing the generalized style was also calculated. Based on the obtained set of vectors, the parameters ε_m ($m = 1, 2, ..., 13$) were calculated, and the authors were ranged according to their ε_m-values. The results of calculations are shown in Figure.

Using the linear regression model, the following dependence was obtained:

$$\varepsilon_m \approx 0.849 - 0.015\ m, \tag{8}$$

i.e. $\varepsilon_0 \approx 0.849$ and $\Delta\varepsilon \approx 0.015$, with the coefficient of determination $R^2 \approx 0.972$. Therefore, the results of our calculations allow us to assert that the parameter ε_m depends linearly on the rank m. However, an important remark should be made here: the linear dependence on the rank was obtained for the selected group of writers. In other words, this empirical law is not universal and cannot be extended onto arbitrary groups of writers. This means that the result is not stable with respect to changes in the group composition. The exclusion of an author from the group (or the inclusion of a new author into the group) may result in a violation of dependence (7). Moreover, the very affiliation or non-affiliation of an author to the group could be characterized by analyzing whether relationship (7) is satisfied.

Another remark concerns the principle of joining the majority *per se*. When formulating it, we considered the vector $\mathbf{a}$ to be constant, which is actually not the case. In effect, we proceeded from the assumption

that each of the vectors $\mathbf{e}_m$ does not affect the vector **a**. It is possible provided that there exists a certain "core" of writers who form the vector of general style, whereas all others join this group. In the limiting case, the "core" can be regarded as consisting of a single, but the most successful, writer. Here, we have an analogy with the mean-field theory that is used in the physics of critical phenomena and phase transitions [48, 49]. It is clear that this case is an approximation, but it does not seem to have a principal value (although it imposes restrictions on the applicability of the results obtained).

5. Conclusions

The model proposed in this work formalizes the process of identifying the authorial style. In particular, the model makes it possible to check the similarity of literary styles of various authors by analyzing various groups or categories of used lexemes. The model allows a quantitative evaluation of the relevant characteristics and not only their qualitative, but also quantitative comparison to be made. The approach used in this work can be useful for the general classification of text stylistics, including the automatic mode.

1. M. Tsizh, B. Novosyadlyj, Yu. Holovatch, N.I. Libeskind. Large-scale structures in the ΛCDM Universe: network analysis and machine learning. *Month. Not. R. Astronom. Soc.* **495**, 1311 (2020).
2. Yu. Holovatch, M. Dudka, V. Blavatska, V. Palchykov, M. Krasnytska, O. Mryglod. Statistical physics of complex systems in the world and in Lviv. *Zh. Fiz. Dosl.* **22**, 2801 (2018) (in Ukrainian).
3. Yu. Holovatch, M. Dudka, V. Blavatska, V. Palchykov, M. Krasnytska, O. Mryglod. *Statistical Physics of Complex Systems.* Preprint ICMP-17-06U (Institute for Condensed Matter Physics, Lviv, 2017) (in Ukrainian).
4. 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. MacCarron, P. MacCarron (Springer, 2016).
5. Yu. Holovatch, R. Kenna, P. MacCarron, P. Sarkanych, N. Fedorak, J. Yose. Mathematics and myths: A quantitative approach to comparative mythology. *Ukr. Modern.* **27**, 108 (2020) (in Ukrainian).
6. R. de Regt, C. von Ferber, Yu. Holovatch, M. Lebovka. Public transportation in UK viewed as a complex network. *Transportmetrica A* **15**, 722, (2019).
7. Yu. Holovatch, R. Kenna, S. Thurner. Complex systems: Physics beyond physics. *Eur. J. Phys.* **38**, 023002 (2017) [arXiv: 1610.01002].
8. B. Berche, C. von Ferber, T. Holovatch, Yu. Holovatch. Transportation network stability: A case study of city transit. *Adv. Compl. Syst.* **15**, 1250063 (2012) [arXiv: 1201.5532].
9. F. Jovanovic, C. Schinckus. Econophysics: A new challenge for financial economics. *J. Hist. Econom. Thought* **35**, 319 (2012).
10. R. Mantegna, H. Stanley. *An Introduction to Econophysics* (Cambridge Univ. Press, 2000).
11. C. Schinckus, F. Jovanovic. Towards a transdisciplinary econophysics. *J. Econom. Method.* **20**, 164 (2013).
12. D. Stauffer. A biased review of sociophysics. *J. Stat. Phys.* **151**, 9 (2013) [arXiv: 1207.6178v1].
13. C. Castellano, S. Fortunato, V. Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591 (2009) [arXiv: 0710.3256].
14. S. Galam. Sociophysics: A review of Galam models. *Int. J. Mod. Phys. C* **19**, 409 (2008)[arXiv: 0803.1800].
15. O.M. Vasilev. A model of rumors spreading in the community with opportunistic behavior. *Zh. Fiz. Dosl.* **22**, 3801 (2018) (in Ukrainian).
16. O.M. Vasilev, O.V. Chalyi. The modeling of macroeconomic dynamics by the methods of econophysics. *Zh. Fiz. Dosl.* **17**, 4801 (2013) (in Ukrainian).
17. A. Rovenchak, S. Buk. Quantum distributions and research of texts: temperature and literature. *Ukr. Modern.* **27**, 29 (2020) (in Ukrainian).
18. S.N. Buk, Yu. Krynytskyi, A. Rovenchak. Properties of autosemantic word networks in Ukrainian texts. *Adv. Compl. Syst.* **22**, 1950016 (2019).
19. A. Rovenchak, S. Buk. Part-of-speech sequences in literary text: Evidence from Ukrainian. *J. Quantit. Linguist.* **25**, 1 (2017).
20. A.A. Rovenchak, S. Buk. Defining thermodynamic parameters for texts from word rank-frequency distributions. *J. Phys. Stud.* **15**, 1005 (2011).
21. A.A. Rovenchak, S. Buk. Application of a quantum ensemble model to linguistic analysis. *Physica A* **390**, 1326 (2011) [arXiv: 1011.5076].
22. Yu. Holovatch, V. Palchykov. Fox Mykyta and networks of language. *Zh. Fiz. Dosl.* **11**, 22 (2007).
23. O. Vasilev, O. Chalyi, I. Vasileva. Mathematical methods and models in linguistics. *Ukr. Modern.* **27**, 9 (2020) (in Ukrainian).
24. A.N. Vasilev, I.V. Vasileva. Physics beyond physics: Application of physical approaches in quantitative linguistics. *Ukr. J. Phys.* **65**, 143 (2020).
25. A. Vasilev, I. Vasileva. Text length and vocabulary size: Case of the Ukrainian writer Ivan Franko. *Glottometrics* **43**, 1 (2018).
26. A.N. Vasilev, A.V. Chalyi, I.V. Vasileva. About "exotic" problems of physics, Winnie the Pooh and Zipf's law. *Zh. Fiz. Dosl.* **17**, 1001 (2013) (in Ukrainian).
27. O.M. Vasiliev, I.V. Vasilieva. Features of the creation of mathematical models in linguistics. *Visn. Kherson. Nats. Tekhn. Univ.* **69**, 99 (2019) (in Ukrainian).

28. Yu. Tuldava. *Problems and Methods of Quantitative-Systemic Research of Lexicon* (Valgus, 1987) (in Russian).
29. R. Piotrovskii, K. Bektaev, A. Piotrovskaya. *Mathematical Linguistics* (Vysshaya Shkola, 1977) (in Russian).
30. V.V. Levitskii. *Quantitative Methods in Linguistics* (Ruta, 2005) (in Russian).
31. N. Darchuk, I. Denysenko, O. Siruk, V. Sorokin. Theoretical issues of modeling the ideographic thesaurus of the Ukrainian language. *Ukr. Movozn.* **24**, 107 (2002) (in Ukrainian).
32. N.P. Darchuk, L.A. Aleksienko, V.M. Sorokin. Parameterized database of poetic speech as a source of philological studies. *Mova Probl. Prykl. Lingvist.* **9**, 15 (2004) (in Ukrainian).
33. N.P. Darchuk. Poetic dictionary from the viewpoint of the world's linguistic picture. *Ukr. Movozn.* **35**, 55 (2006) (in Ukrainian).
34. N.P. Darchuk. Research corpus of the Ukrainian language: Basic principles and prospects. *Visn. Kyiv. Nats. Univ. Literat. Lingvist. Folklor.* **21**, 45, (2010) (in Ukrainian).
35. N.P. Darchuk. Automatic syntactic analysis of the texts in the corpus of the Ukrainian language. *Ukr. Movozn.* **43**, 11 (2013) (in Ukrainian).
36. N.P. Darchuk. Semantics formalization directions. *Movn. Kontsept. Kart. Svit.* **46**, 385 (2013) (in Ukrainian).
37. R.T. Grom'yak, Yu.I. Kovaliv, V.I. Teremko. *Literary Dictionary-Reference Book* (Akademiya, 1997) (in Ukrainian).
38. H.S. Green, C.A. Hurst. *Order-Disorder Phenomena* (Interscience, 1964).
39. B.M. McCoy, T.T. Wu. *The Two-Dimensional Ising Model* (Cambridge Univ. Press, 1973).
40. G. Salton, A. Wong, C.S. Yang. A vector space model for information retrieval. *Commun. ACM* **18**, 613 (1975).
41. W. Pauli. *General Principles of Quantum Mechanics* (Springer, 1980).
42. A.S. Davydov. *Quantum Mechanics* (Pergamon Press, 1976).
43. Th. Veblen. *The Theory of the Leisure Class* (Courier Corporation, 1994).
44. H.A. Simon. A Behavioral Model of Rational Choice. *Quart. J. Econom.* **69**, 99 (1955).
45. N.P. Darchuk. Structural-statistical database of the modern Ukrainian language on the basis of frequency dictionaries. In *Vocabulum et Vocabularium* (Grognen. Gos. Univ., 2005) (in Russian).
46. L.A. Alekseenko, N.P. Darchuk, O.N. Zuban', V.V. Sorokin. Parameterized database of poetic speech as a source and an instrument for philological studies. In: *Proceedings of the International Conference "Computer Linguistics without Borders"* (St. Petersburg, 2004).
47. I. Vasileva. The application of computer thesaurus in the study of the poets' language. *Leksykogr. Byulet.* **13**, 161 (2006) (in Ukrainian).
48. Shang-Keng Ma. *Modern Theory of Critical Phenomena* (Benjamin, 1976).
49. A.Z. Patashinskii, V.L. Pokrovskii. *Fluctuation Theory of Phase Transitions* (Pergamon Press, 1982).

Received 02.06.20.
Translated from Ukrainian by O.I. Voitenko

Н.П. Дарчук, І.В. Васильєва, О.М. Васильєв

ВЕКТОРНА МОДЕЛЬ
АНАЛІЗУ СТИЛІСТИКИ ТЕКСТІВ

Стаття присвячена застосуванню фізичних підходів до аналізу авторських стилів українських письменників. Пропонується модель, у якій літературні стилі описуються векторами одиничної довжини в багатовимірному просторі. Числовою характеристикою стилю є результат скалярного добутку відповідного вектора на вектор, який визначає загальний стиль для групи авторів. Показано, що цей параметр лінійним чином залежить від рангу автора. Така залежність підтверджує гіпотезу приєднання до більшості, відповідно до якої автори, вибираючи стиль, орієнтуються на стиль своїх успішних колег.

Ключові слова: фізика складних систем, квантитативна лінгвістика, вектор, енергія, розподіл.