On some concepts of nonlinear dynamics suitable for use in linguistics

A A Kretov\textsuperscript{1}, M V Polovinkina\textsuperscript{2}, I P Polovinkin\textsuperscript{3} and M V Lometc\textsuperscript{1}

\textsuperscript{1}Department of Theoretical and Applied Linguistics, Voronezh State University, 10 pl. Lenina, Voronezh, 394006, Russian Federation
\textsuperscript{2}Department of higher mathematics and information technologies, Voronezh State University of Engineering Technologies, 19, Revolution Av., Voronezh, Russian Federation
\textsuperscript{3}Department of mathematical and applied analysis, Voronezh State University, 1, Universitetskaya pl., Voronezh, Russian Federation

E-mail: polovinkin@yandex.ru

Abstract. In this paper, we describe hypothetical possibilities for applying nonlinear dynamics methods to linguistic research. We suggest to refer to a large text corpus (metabook of the writer, national language, language of the branch of knowledge) as a fractal object and to measure its fractal dimension (by Hausdorff). We consider the question of practical calculation of fractal dimension, as well as the question of motivation for applying the fractal concept to the language. In addition, we propose to apply some methods of the qualitative theory of differential equations to modelling the growth of the natural language vocabulary.

1. Introduction

The non-linear dynamics theory is faced with many phenomena in various fields of activity, where its apparatus can be applied. In particular, this can be said about the principle of self-similarity and the concept of a fractal. We can say that the concept of a fractal in mathematics and physics has taken shape. In other fields, the detection of self-similarity effects and the ability to use the tools of the fractal theory are rather scattered, although the base of the revealed facts is quite extensive. In this paper, we propose to look at some of the achievements in modern linguistics from the fractal theory point of view. Fractal (self-similar) manifestations in the language have been noticed in linguistics researching (see, for example, [1–3]). Basically, we are talking about the statement and verbal description of self-similarity in the language. However, there is every reason to consider the quantitative characteristics of the language fractality.

Another area where the ideology of nonlinear dynamics can be used is related to the study of the formation and change of language over time. The information on the quantitative characteristics of the language change process, accumulated to date, suggests that this process proceeds according to laws similar to the laws of population growth and distribution. This observation makes it possible to apply mathematical methods, already successfully used in biology, medicine and even in economics, for studying the language development.

2. Self-similarity and fractality of language

We assume that the language as an object of research has (or at least can be endowed by the researcher) one of the main properties that distinguish the fractal, namely the property of self-
similarity. To illustrate one of the manifestations of self-similarity, let us turn to the concept of the core and periphery of the language (for more information, see the monograph [4], according to which we construct the narration in this section).

The concepts of the core (kernel, center) and periphery are firmly established in linguistics [5]. There are at least two understandings of the core. In one of them, the most typical mass-like phenomena belong to the core. So, there are more single-valued words in the vocabulary than two-valued or three-valued ones, therefore, on this basis, single-valued words should be attributed to the core, first of all, then two-valued, three-valued and more multi-valued words will be assigned to the periphery.

In the second approach, the core includes those phenomena that have the maximum degree of manifestation of a trait, and the periphery includes the phenomena with the minimum degree of manifestation of the trait. In this case, the core will include words with the maximum number of values: 20, 15, 10, and the periphery will include one-, two-, three-valued words.

These two approaches and types of cores should be different terminologically. For example, in the first case, we can talk about a quantitative core, and in the second case, we can talk about a qualitative core. In our case, we are talking about qualitative cores, but as far as the language is concerned, the qualitative core is a quantitative periphery and Vice versa, therefore, no matter what core and periphery are meant (quantitative or qualitative), the boundary between them will remain unchanged and in this sense, it will be universal.

The answer to the question about the natural boundary between the frequency core and the periphery should be found in the analysis of specific texts and frequency dictionaries compiled on their basis.

It is known that the first thousand of the most common words of the Russian language corresponds to the coverage of 60 to 80% of the text. If we take the next 2000 words, the text coverage will increase by no more than 20%. If we take the next 4th-5th thousand, they will cover less than 10% of the text; the 6th-7th thousand will cover about 5% of the test, and with each subsequent interval of 2000 words, the percentage of text coverage will decrease. In the beginning, the text growth (in %) significantly outstrips the vocabulary growth, and at the end, the vocabulary growth significantly outstrips the text growth. Therefore, there must be a point at which there is a change of laws, a point at which the relative growth of the vocabulary begins to overtake the relative growth of the text, and we, increasing the vocabulary, stop acting rationally, spending more effort for the sake of a smaller result. This point will be the natural boundary between the core and the periphery of the frequency dictionary.

In the first approximation, that point can be found as the total length of the text \( M \) (in word usages) divided by the number of different words \( N \) (word forms, linguistic units, lemmas). This is the average frequency of a word \( F_{av} \) (or word form) in the text in question:

\[
F_{av} = \frac{M}{N},
\]

(1)

\[
F_P < F_{av} < F_C,
\]

(2)

where \( F_P \) and \( F_C \) are frequencies of words in the periphery and in the core correspondingly.

Here we should refer to A. I. Kuznetsova’s idea [6] that the language core is arranged on the principle of matryoshka (Russian nested doll). Let’s call the core of the whole language the first. The first core itself has its own core and periphery, which can be detected if we use the parameters of the first core. For example, for Russian language, taking 4.348 words for 100% of the vocabulary \( (N_2) \), and 902.725 word forms for 100% of the whole text \( (M_2) \) and applying the same procedure to them as to the vocabulary as a whole, we get \( F_{av2} = 207.6 \). Thus, the boundary between the second core and the second periphery is found. Let us repeat this procedure until we get the single most frequently used word. The Russian language core "matryoshka" [7] consists of 7 cores. These data are comparable to data from other frequency
dictionaries (see [4] for details). This applies to the frequency dictionaries of different languages, as well as to the frequency dictionaries of individual works and meta-books of writers, etc. At the same time, in all considered examples, the number of cores not go beyond 7+2 out of dependency on fluctuations in the size of frequency dictionaries. An interesting picture is revealed when replacing the absolute characteristics of the cores with relative ones. We take as a unit the fraction of text covered by the smallest core equal to the most frequent word. Then it turns out [4] that the relative values of the cores in the frequency vocabulary are determined by a sequence of numbers, which is set using recurrent relations of the Fibonacci type. Of course, this is a statistical statement. Its validity was confirmed using the standard procedure for testing the statistical hypothesis. The above means that the structure of the language and text is endowed with the properties of self-similarity and recursiveness, that is, it can be classified as a fractal.

3. Fractal dimension of the meta-book text

In [8], the authors made an attempt to clarify Heaps law (with reference to [9]), according to which a number of different (unique) words (N) in a book, as a function of the total number of words in the book (M), has an order of growth $\Theta (M^\alpha)$ with $\alpha \in (0, 1)$. Considering the Heaps law not as an asymptotic estimate, but as an exact formula with a variable exponent, the authors rewrite it as

$$\alpha = \alpha(M) = \frac{\ln N}{\ln M}. \quad (3)$$

We consider this approach as a reason to turn to the apparatus developed in the theory of fractals. More precisely we regard this formula as a step to calculate the fractal dimension of a text.

In the book [10], Benoit Mandelbrot describes the following approach to the concept of fractal dimension (see also [11]). Let’s choose a set of congruent ”atomic” sets in space $R^d$ that have topological dimension $d$. This is a set of either $d$-dimensional balls or $d$-dimensional cubes. For certainty, we assume that these are balls. Suppose that the fractal object belongs to space $R^d$. Let’s fix a sufficiently small radius $l$ and cover the entire fractal object with balls of radius $l$. Assume that this procedure requires at least $N$ balls. The number

$$\alpha_0 = -\lim_{l \to 0} (\ln N/\ln l) = \lim_{l \to 0}(\ln N/\ln(1/l)) \quad (4)$$

is called the fractal dimension of the object in question.

In form (4), this definition is hardly suitable for describing the text, since we can not make the size of the atomic set, which is naturally considered to be a word (word usage), tending to zero. We have to change it a little in order to adapt it to our needs. In the denotations by [8], we put

$$l = 1/M. \quad (5)$$

It is possible to interpret equality (5) as follows. Considering every word usage as an ”atomic brick” for a text in question, we determine its size by comparing this ”brick” with the text itself, since, in fact, there is nothing else to measure it with. In other words, for the size of the ”atom” we take the share occupied by it as a whole. By the power of text coverage, we mean the number of unique words (lemmas) whose word usage made up the entire text. Next, by definition, we put

$$\alpha_0 = -\lim_{l \to 0} (\ln N/\ln l) = \lim_{M \to +\infty}(\ln N/\ln M) = \lim_{M \to +\infty} \alpha(M), \quad (6)$$

and the number defined by formula (6) is called the fractal dimension of the language. Here, by language, we mean both the national language and the language of the writer or the language of the branch of knowledge (for example, the language of scientific literature in chemistry, mathematics, the language of journalism, etc.).
4. On the practical calculation of the fractal dimension of a text

Formula (6) assumes that the text volume, which is understood as the number of words used in it, can take arbitrarily large values. If we are talking about the text of a certain work, then, of course, this is not the case. The authors of the work [8] mentioned above, introduce the concept of a writer’s meta-book as an association of all texts written by this writer. If a writer is sufficiently prolific, then this concept allows us to assume that \( M \to +\infty \). Although, in practical calculation, we still have to confine the available length of the meta-book for calculating the approximate value \( \alpha_0 \). In [8], the authors claim and illustrate with examples of texts by three different authors (Hardy, Melville and Lawrence) that \( \alpha \) decreases as \( M \) increases. Our observations of Tolstoy’s texts do not refute this conclusion. For our purposes, the Leo Tolstoy meta-book is a very convenient object of research, since Leo Tolstoy owns both relatively small and very voluminous (“War and peace”) works, and the range between the lengths of small and large works is filled quite tightly.

We consider it possible to start from the position that \( \alpha \) is a decreasing function of the variable \( M \). As is known, the limit of a function decreasing on an interval at the right end of this interval is equal to the infimum of this function on the interval. Therefore, the value of \( \alpha \) at the maximum value in the specified range should be considered as the best approximation for an upper bound of the fractal dimension. The lower bound for the fractal dimension of a meta-book can be obtained from the following considerations. On the basis of empirical data, we approximate the function that expresses the dependence of the vocabulary value on the meta-book value. Using the obtained dependence, with the help of extrapolation, we will determine such a value of the meta-book, exceeding which the increment of the vocabulary value will be negligible. Let’s find the corresponding limiting volume of the vocabulary and calculate the value (3) for the found values.

A slightly modified approach may be as follows. Let’s turn to an important characteristic of the meta-book called ”lexical diversity” (LD). LD is a quantitative characteristic of the text that reflects the degree of richness of the vocabulary when constructing a text of a given length. In the simplest version, LD is calculated as the ratio of the number of individual lexical units of the vocabulary (types) to the number of their uses in the text (tokens). For this method of calculation, the notation TTR (type/token ratio) is accepted. TTR was presumably introduced into scientific use in 1957 by a specialist in linguodidactics M. Templin [12]. The calculation of LD as TTR is criticized for not taking into account the effect of text length, because as the text length increases, the vocabulary size grows more slowly, and TTR will decrease and tend to zero. However, it is the quality of TTR that is useful for our purposes. It is possible to consider the maximum vocabulary size to be such a value of this size, at which LD becomes negligible. In this regard, it is necessary to clarify what is meant by ”smallness” both of the vocabulary increment and of LD. Here there is a problem to link this concept of smallness with the choice of a trend model and, as a consequence, the method of extrapolating the trend. As an example of the application of the above considerations, we examined 20 works of Leo Tolstoy of different volumes, covering more or less evenly a time span of 52 years. At the same time, we deliberately took texts of different sizes to deal with the most difficult case of new word growth. We had to take 19 steps, at each of these steps we expanded the meta-book by concatenating the text of the next work, we calculated its current size equal to the number of tokens, and we also carried out lemmatization, the corresponding expansion of the vocabulary and the calculation of its current size. We carried out lemmatization using the publicly available morphological analyzer of the Russian language MyStem developed by Ilya Segalovich at Yandex. On the basis of calculations made with this meta-book, we came to an upper estimate of the fractal dimension of L.N. Tolstoy equal to 0.7252.

Let’s choose a logarithmic dependence as a trend line. More precisely, we choose logarithmic and constant functions as the basis, and we look for the function of the dependence of LD on
Table 1. The dynamics of LD in the growing corpus of texts.

| No | Year | Title                        | Tokens qt | Types qt | Meta–book size | Vocabulary size | TTR    |
|----|------|------------------------------|-----------|----------|----------------|-----------------|--------|
| 1  | 1852 | Childhood                   | 30326     | 4253     | 30326          | 4253            | 0,1402 |
| 2  | 1854 | Adolescence                 | 23020     | 1452     | 53346          | 5705            | 0,1069 |
| 3  | 1855 | Sevastopol stories          | 36041     | 2117     | 89387          | 7822            | 0,0875 |
| 4  | 1856 | Two husars                  | 17219     | 844      | 106606         | 8666            | 0,0813 |
| 5  | 1856 | Morning of the landowner    | 15669     | 751      | 122275         | 9417            | 0,0770 |
| 6  | 1857 | Youth                       | 49939     | 1208     | 172214         | 10625           | 0,0617 |
| 7  | 1858 | Albert                      | 7927      | 147      | 180141         | 10772           | 0,0598 |
| 8  | 1862 | Polikushka                  | 16879     | 725      | 197020         | 11497           | 0,0584 |
| 9  | 1863 | The Cossacks                | 46002     | 1391     | 243022         | 12888           | 0,0530 |
| 10 | 1869 | War and Peace               | 459672    | 7212     | 702694         | 20100           | 0,0286 |
| 11 | 1877 | Anna Karenina               | 270110    | 2702     | 972804         | 22802           | 0,0234 |
| 12 | 1884 | Diary of a Madman           | 3729      | 32       | 976533         | 22834           | 0,0234 |
| 13 | 1886 | Death of Ivan Ilyich        | 17716     | 190      | 994249         | 23024           | 0,0232 |
| 14 | 1889 | The Kreutzer Sonata         | 25434     | 270      | 1019683        | 23294           | 0,0228 |
| 15 | 1890 | Devil                       | 14246     | 395      | 1033929        | 23689           | 0,0229 |
| 16 | 1891 | Mother                      | 3597      | 48       | 1037526        | 23737           | 0,0229 |
| 17 | 1895 | Master and Man              | 14270     | 268      | 1051796        | 24005           | 0,0228 |
| 18 | 1898 | Father Sergius              | 13706     | 153      | 1065502        | 24158           | 0,0227 |
| 19 | 1899 | The Awakening               | 137305    | 1591     | 1202807        | 25749           | 0,0214 |
| 20 | 1904 | Hadzhi–Murat                | 36376     | 481      | 1239183        | 26230           | 0,0212 |

The volume of the text in the form of a linear combination of basic functions. The likelihood ratio in this case is also very high ($R^2 \approx 0,9611$). We obtained

$$Y_{TTR} = 0,4081 - 0,028 \ln X_{text},$$

(7)

where $X_{text}$ is the size of the current meta-book, $Y_{TTR}$ is the current value of TTR. The value of the text size at zero of this function can be considered as corresponding to the maximum
vocabulary size. This function reaches zero at the point

\[ X_{\text{text}} - 0 \approx \exp(14.57) \approx 2 \, 129 \, 565. \]  

(8)

Thus, based on the chosen method of modelling, we conclude that the size of the meta-book, at which the maximum size of Leo Tolstoy’s vocabulary is achieved, equals 2 129 565 words. It is clear that this is some approximate estimate. We will find the maximum vocabulary size from the same considerations with the same choice of basis functions.

\[ Y_{TTR} = 0.6301 - 0.0604 \ln X_{\text{voc}}. \]  

(9)

where \( X_{\text{voc}} \) is a current size of the vocabulary. This function reaches zero at the point

\[ X_{\text{voc}} - 0 \approx \exp(10.43) \approx 33 \, 932. \]  

(10)

Therefore, the estimate of the maximum size of L. N. Tolstoy’s vocabulary (with the necessary remark about taking into account the chosen modelling method) is approximately 33,932 words.

There is one more problem. We are referring to the problem of verifying the validity of the obtained forecasts. The classical method of comparing an approximate solution with an exact solution or with experimental data cannot be applied due to the lack of such data. Here we can only use indirect methods of verification. Nevertheless, we will try to use a variant of Zipf’s law

\[ X_{\text{voc}} = AX_{\text{text}}^\gamma \]  

for describing the dependence of the vocabulary size on the text size. This is justified by the fact that such a dependence will be obtained if from the equations (7)–(9) we will exclude \( Y_{TTR} \). This way of expressing \( X_{\text{voc}} \) through \( X_{\text{text}} \) will result in the approximate formula

\[ X_{\text{voc}} = 39,5356X_{\text{text}}^{0.4636} \]  

(11)

Substituting (8) in formula (11), we obtain

\[ X_{\text{voc}} - 0 \approx 33 \, 945, \]  

(12)

which is quite good. But this, of course, cannot be considered a method of verification. It is another matter if we use the primary data from the table and approximate the power function in Zipf’s law. This approach gives us

\[ X_{\text{voc}} = 38,069X_{\text{text}}^{0.4653} \]  

(13)

Now, substituting the value \( X_{\text{text}} = 2 \, 129 \, 565 \), we obtain

\[ X_{\text{voc}} - 0 \approx 33 \, 506. \]  

(14)

This value differs from the one obtained earlier as the zero of the logarithmic function of the LD trend. However, the relative error is

\[ \frac{33 \, 932 - 33 \, 506}{33 \, 506} \times 100\% \approx 1.27\%, \]

which seems, in our view, to be quite acceptable. It remains only to make the final decision on the forecast of the maximum vocabulary size and the corresponding size of the meta-book. Having made traditional rounding, we come to the following forecasts: the maximum size of L. N. Tolstoy’s vocabulary is 33 500–34 000 words, the text size providing the maximum size of L. N. Tolstoy’s vocabulary is 2129.500–2130 000 words. Now we can calculate the lower bound of the fractal dimension:

\[ \alpha_0 \approx \frac{\ln 34 \, 000}{\ln 2130 \, 000} \approx 0,7160. \]

Thus, the fractal dimension of L. N. Tolstoy’s meta-book composed of his 20 books, can be enclosed in the interval \([0.7160; 0.7252]\).
5. On modelling the growth of a natural language vocabulary

Let us now turn to the question of modelling the growth of a language vocabulary. Very informative thoughts on this problem can be found in books [13] and [14]. We will make a brief introduction to this section, based on these books (unfortunately, each of them is a bibliographic rarity).

It is a well-known thesis that “as a result of the constant expansion of the sphere of human activity, the vocabulary of each language, especially its terminological dictionary, despite the loss of a certain number of words, is steadily growing” ([13], p. 56). Such a steady increase in the size of the dictionary is associated, in particular, with an exponential growth law according to the following formula ([13], p. 57):

\[ L(t) = L_0 e^{nt}, \]  

where \( t \) is the point in time, \( L(t) \) is the vocabulary size at time \( t \), \( L_0 \) is the initial size of the dictionary, which specifies the start of the time frame, \( n > 0 \) is the growth rate. According to the exponential law, the growth rate of the vocabulary has an "avalanche" character (the growth rate of the vocabulary proportional to the achieved level), which can be described by the ordinary differential equation

\[ \frac{dL(t)}{dt} = nL(t). \]  

The factor \( L_0 \) is specified by the initial condition

\[ L(0) = L_0. \]  

A retrospective check of the law (15) based on data on representative (for example, explanatory) dictionaries of some natural languages proves that the growth of common vocabulary and the growth of the volume of the dictionary of a literary language over different periods of time can be characterized by an exponential law only in certain periods of language development [14]. In fact, the process of lexical growth begins slowly (the period of formation of a literary language), then accelerates and takes on an "avalanche" character (the period of formation of a literary language), but at some point the growth process necessarily slows down (the period of stabilization). Such a scheme of development corresponds to a mathematical model, expressed by the so-called logistic function

\[ L(t) = \frac{L_c}{1 + ae^{-kt}}, \]  

where \( L_c \) is the theoretical supremum of the vocabulary, \( k > 0, a \) are some parameters. Graphically, this model is represented by an S-shaped curve, which expresses the downloaded growth with increasing speed, then the speed decreases and almost stops as the asymptotic approach to the limit \( L_c \).

In all probability, the law of logistic development in its General form (acceleration — inflection point — deceleration) has a General socio-linguistic significance and characterizes the growth and development of the vocabulary of most literary languages, although this law takes a specific form depending on the conditions of the historical development of a given people — a native speaker. It may be added that the logistic law of growth in its various concrete manifestations (there are a number of variant formulas of logistic growth) is considered one of the main laws of development of self-organizing complex systems, if we consider their development at sufficiently large time intervals. Some other diachronic linguistic processes are also characterized by an S-curve. They are now widely used in many areas of science, including in solving problems of modelling the development of science itself (see [14], p. 155). It should be mentioned that the
vocabulary growth modelling develops in the same way as the population growth modelling with the formulation of similar quantitative patterns.

Equation (16) is called the Malthusian model (see [15]). One of the first researchers into population dynamics was Thomas Malthus. Malthus observed in an essay written in 1798 that the growth of the human population was fundamentally different from the growth of the food supply to feed that population. He wrote that the human population was growing geometrically (i.e. exponentially) while the food supply was growing arithmetically (i.e. linearly). He concluded that left unchecked, it would only be a matter of time before the world’s population would be too large to feed itself. Malthus assumed that the rate, at which the population grew, was directly proportional to its current size. If the population at time \( t \) is denoted by \( L(t) \), then the assumption of natural growth can be written symbolically as (16) and \( L_0 \) is the initial population. The solution predicts

- population explosion if \( n > 0 \),
- population extinction if \( n < 0 \),
- no change if \( n = 0 \).

The Malthusian model is commonly called the natural growth model or the exponential growth model. This model may be useful in situations in which the time scale of observation is small enough to make it acceptable to assume that \( n > 0 \) remains nearly constant, resources appear to be unlimited, and \( L_0 \) is small.

Later Pierre François Verhulst (1838) replaced the constant relative growth rate \( n \) in (16) by the relative growth rate

\[
k \left( 1 - \frac{L}{L_c} \right),
\]

that decreases linearly as a function of \( L \). The dimensionless factor \( k(1-L/L_c) \) serves to diminish the relative growth rate from \( k \) down to zero as the population increases from its initial level \( L_0 \) to \( L_c \). The constant \( L_c \) represents the maximum sustainable population beyond which \( L \) cannot increase. The resulting model (see [15]),

\[
\frac{dL}{dt} = kL \left( 1 - \frac{L}{L_c} \right), \tag{19}
\]

is called the logistic growth model or the Verhulst model. The Verhulst model assumes that the growth rate declines from a value \( k \), when conditions are very favorable, to the value 0, when the population has increased to the maximum value \( L_c \) that the environment can support. The solution of equation (19) is

\[
L(t) = \frac{L_cL_0}{L_0 + (L_c - L_0)e^{-kt}},
\]

and it coincides with (18) with \( a = (L_c - L_0)/L_0 \).

The logistic model predicts rapid initial growth for \( 0 < L_0 < L_c \), then a decrease in growth rate as time passes so that the size of the population approaches a limit. This behavior is in agreement with the observed behavior of many populations, and for this reason, the logistic model is often used as a means of describing population size.

The Verhulst model and the Malthusian model didn’t take into account population migration. Harold Hotelling (1921) added to the Verhulst equation diffusion term describing migration. As a result, the equation took the form (see [16])

\[
\frac{\partial L}{\partial t} = A(L_c - L)L + B\Delta L, \tag{20}
\]

where \( L = L(x_1, x_2, t) \) is the population density at point \((x_1, x_2)\) at time \( t \),

\[
\Delta = \nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}
\]
is the Laplace operator,

\[ \nabla = \left( \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_1} \right) \]

is the Hamilton operator (symbolic vector), \( A \) is the population growth rate, \( B \) is the migration rate, \( s \) is the coefficient of the saturated population density, \( t \) is the time parameter, \( x_1, x_2 \) are the geographical coordinates.

The Hotelling model describes both population growth and migration processes. Population growth is modelled as a logistic process. Migration processes are described using Fourier’s Law of Heat Conduction. Hotelling introduced the notion of saturated population density. If a real population density is higher than a saturated one, the population decreases, if a real population density is lower than a saturated one, the population increases. The spatial diffusion was explained by the fact that population growth declines output per capita, and people move from more populated places to less populated ones.

A significant weakness of the Hotelling model is that livelihood stocks are assumed to be equal to a given constant, not depending on time and population (labour force). Therefore, this model is more suitable for animal populations, as evidenced through its successful application in ecology in 30 years after creation.

In a large number of cases, when trying to build a model of an object, it is either impossible to directly specify the fundamental laws or variational principles to which it obeys, or, from the point of view of our current knowledge, there is no confidence at all in the existence of such laws that allow a mathematical formulation. One of the most fruitful approaches to such objects is the use of analogies with already studied phenomena (see [17]). In the case of the vocabulary growth modelling, we can repeat fragments of the research path for population growth and distribution. Often, along with the explicit definition of functions, which is not always known, it is useful to study the differential equations that this function satisfies, developing a law that describes this equation. In our case, along with specifying function (18), we can explicitly consider the differential equation (19) that sets it. This approach is useful, for example, in the study of issues related to the stability of the stationary state. Equation (19) is well studied. We can go further along the same path as in the study of population growth and distribution. We add a diffusion term to the right of the equal sign and get an equation of the form (20), where \( L = L(x_1, x_2, t) \) is the vocabulary size at point \((x_1, x_2)\) at time \(t\). In some cases, it is convenient to interpret \( L = L(x_1, x_2, t) \) as the deviation of the vocabulary size at time \(t\) from the size fixed as a certain stationary level, which is assumed to be zero. Unfortunately, such a ”localized” approach can lead to additional technical difficulties in experimental confirmation of the model. Without the diffusion term \( B\Delta L \), we are dealing with a ”point” model, which is actually just global in the sense that law (20) can be applied to the language as a whole. In this case, one can rely on representative dictionaries of the language under consideration from different years to test hypotheses. In the case of diffusion-logistic model (20), we may, if we are talking about a large country, force ourselves to replace the study of the national language with the study of adverbs and dialects, the dynamics of which are not so detailed in official dictionaries.

However, we can hope that this model will be suitable for studying the influence of the interpenetration of languages of a certain territory with relatively small subterritories (countries, federal lands, regions, etc.) whose languages and adverbs are recorded in dictionaries. The diffusion approach actually provides new knowledge about the model. The simplest example illustrating the above is the problem of the stability of a stationary state. Let \( w = w(x_1, x_2) \) be a stationary solution of equation (20) in some bounded domain \( \Omega \) with a piecewise smooth boundary, satisfying on the boundary of this domain the classical boundary condition of the first, second, or third kind, \( d \) is a diameter of the domain \( \Omega \).
In [18], it is shown that the condition

\[ w > \frac{L_c}{2} - \frac{B}{2Ad^2} \]

is sufficient for the stability of the stationary solution \( w(x_1, x_2) \) (see also [19], where this result is generalized). It is well known that without a diffusion term (for \( B = 0 \)), the zero stationary solution is unstable. In the diffusion case (for \( B \neq 0 \)), the trivial stationary solution can be both stable and unstable, which is determined by the size of the domain \( \Omega \). This is exclusively the effect of the diffusion model. In our interpretation, this means the following qualitative conclusion: migration processes contribute to keeping the size of the vocabulary in a stable state, when it is the vocabulary of a language used in a relatively small area. Additional research is required to confirm or refute this conclusion and the adequacy of the model.

References

[1] Kretov A A and Voronina I V 1995 The Russian word as a self-similar recursive structure (in Russian) In: Linguistics at the end of the XX century: results and prospects: collection of scientific works 1 Moscow, Philology 269–71
[2] Bronnik L V 2009 On fractal self-similarity in language (in Russian) Proceedings of the Volgograd state pedagogical University 44 10 15–9
[3] Petryakov L D 2017 Methodological perspectives of fractal semantics (in Russian) News of universities. Series "Humanitarian Sciences 8 2 148–53
[4] Kretov A A 2006 Basics of lexical and semantic prognostics (in Russian) Voronezh, Publishing house of Voronezh state University
[5] Kuznetsova A I 1988 Parametric study of peripheral phenomena in the field of morphemics: on the material of the Russian language (in Russian) Abstract of the dissertation of the doctor of Philology Moscow.
[6] Kuznetsova A I 1981 Correlation of the center and periphery in the field of morphemics of the Russian language (in Russian) In Scientific and technical conference "Problems of derivatology": Thesis of reports 2 Perm 153–6
[7] Zasorina L N editor 1977 Frequency dictionary of the Russian language: About 40,000 words (in Russian) Moscow, Russian language
[8] Bernhardtson S, Correa da Rocha L E and Minnhagen P 2009 The meta book and size-dependent properties of written language New Journal of Physics 11:123015 Online at http://www.njp.org/ doi:10.1088/1367-2630/11/12/123015.
[9] Heaps H S 1978 Information Retrieval: Computational and Theoretical Aspects New York, Academic Press
[10] Mandelbrot B B 1982 The Fractal Geometry of Nature San Francisco, W H Freeman
[11] Maslov V P 2005 On a General Theorem of Set Theory Leading to the Gibbs, Bose-Einstein, and Pareto Distributions as well as to the Zipf-Mandelbrot Law for the Stock Market Mathematical Notes 78 807–13
[12] Torruella J and Capsada R 2013 Lexical Statistics and Tipological Structures: A Measure of Lexical Richness Procedia - Social and Behavioral Sciences 95 447–54
[13] Piotrowski R G, Bektaev K B and Piotrowskaja A A 1985 Mathematische Linguistik Brockmeyer, Bochum
[14] Tulldava J 1987 Problems and Methods of the Quantitative-Sistemic Investigation of Vocabulary (in Russian) Tallin, Valgus
[15] Brauer F and Castillo-Chavez C 2012 Mathematical Models in Population Biology and Epidemiology New York, Springer; Dordrecht Heidelberg, London
[16] Pun T 1997 Nonlinear economic dynamics Berlin; Heidelberg; New York; Barcelona; Budapest; Hong Kong; London; Milan; Paris; Santa Clara; Singapore; Tokyo, Springer-Verlag
[17] Samarskii A A and Mikhailov A P 2002 Mathematical Modelling: Ideas. Methods. Examples (in Russian) Moscow, Physical and Mathematical Literature Publishing Company Russian Academy of Sciences
[18] Meshkov V Z, Polovinkin I P and Semenov M E 2002 On the stability of a stationary solution of the Hotelling equation (in Russian) Appl. and Industrial Math. Rev. 9 1 226–7
[19] Gogoleva T N, Shchepina I N, Polovinkina M V and Rabeekh S A On stability of a stationary solution to the Hotelling migration equation 2019 J. Phys.: Conf. Ser. 1203 012041