УДК 81`32 ***doi: 10.26907/2658-3321.2020.3.2.71-81***

МЕТОДЫ ПРИМЕНЕНИЯ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ В ЛИНГВИСТИКЕ: ОПЫТ ТЕОРЕТИЧЕСКОГО ИССЛЕДОВАНИЯ

Н.А. Сигачева, А.Р. Баранова

nsigacheva@mail.ru

Казанский (Приволжский) федеральный университет, г. Казань, Россия

Аннотация. Актуальность данной статьи обусловлена появлением новых областей языкознания, находящихся «на стыке» с естественнонаучной сферой, проникновением в лингвистику математических методов способствующих развитию языкознания в сторону точности и объективности, что вызываетнеобходимость дополнительных исследований. В работе применялись как теоретические, так и эмпирические методы научного исследования. Данная статья направлена выявлениеи анализ связи между лингвистикой и математикой. В результате были исследованы примеры применения математической формулы в лингвистике и рассчитаны возможные периоды дивергенции ряда языков. Материалы статьи могут быть полезны как для преподавателей в сфере лингвистики, так и для студентов-филологов.

Ключевые слова: лингвистика, модель, глоттохронология, лексикостатистика, математические методы, терминология.

Для цитирования: Сигачева Н.А., Баранова А.Р. Применение математических моделей в лингвистике: опыт теоретического исследования. *Казанский лингвистический журнал*. 2020; 2 (3): 71–81. (на англ.) DOI: 10.26907/2658-3321.2020.3.2.71–81.

METHODS OF USING MATHEMATICAL MODELS IN LINGUISTICS: THE EXPERIENCE OF THEORETICAL RESEARCH

N.A. Sigacheva, A.R. Baranova

nsigacheva@mail.ru

Kazan Federal University, Kazan, Russia

Abstract. The relevance of this article is due to the emergence of new areas of linguistics that are "at the junction" with the natural science sphere and to the penetration of mathematical

methods in linguistics that contribute to the development of linguistics in the direction of accuracy and objectivity. This calls for additional research. The authors use both theoretical and empirical methods of scientific research. This article aims to identify and analyze the relationship between linguistics and mathematics. As a result of the research, examples of the application of mathematical formula in linguistics were investigatedand possible time of some languages divergence was calculated. The materials of the article can be useful both for teachers in the field of linguistics and for students of Philology.

Keywords: linguistics, model, glottochronology, lexicostatistics, mathematical methods, terminology.

For citation: Sigacheva N.A., Baranova A.R. Methods of using mathematical models in linguistics: the experience of theoretical research. Kazan linguistic journal. 2020; 2 (3): 71–81. DOI: 10.26907/2658-3321.2020.3.2.71–81.

The relevance of the chosen research direction is due to the emergence of more and more new areas of mental activity that are “at the intersection” with the natural science, technical and humanitarian fields. Penetration of mathematical methods into the linguistics contributed to the development of linguistics in the direction of accuracy and objectivity.In the nineteenth century, mathematicians began to develop non-quantitative abstract models that were later used in linguistics. The first attempts to use real mathematical tools to describe linguistic phenomena were undertaken only in the middle of the twentieth century. One of the results of these studies was the emergence of a new mathematical discipline-mathematical linguistics, the subject of which is the development of a mathematical apparatus for linguistic research.

Mathematical linguistics is a mathematical discipline, the subject of which is the development of a formal apparatus for describing the structure of natural and some artificial languages. It appeared in the 50s of the 20th century. One of the main incentives for the emergence of mathematical linguistics was the need to clarify its basic concepts in linguistics. The methods of mathematical linguistics have much in common with the methods of mathematical logic – a mathematical discipline that studies the structure of mathematical reasoning – and especially its sections such as the theory of algorithms. Algebraic methods are also widely used in mathematical linguistics. Mathematical linguistics is developed in close

cooperation with linguistics. Sometimes the term "*mathematical linguistics*" is also used to refer to any linguistic research that uses a mathematical apparatus.

The mathematical description of a language is based on the idea of language as a mechanism that functions in the speech activity of its speakers, which goes back to F. de Saussure. Its result is "correct texts" – sequences of speech units that follow certain laws, many of which allow mathematical description. The development and study of ways of mathematically correct description of the correct texts (primarily sentences) is the content of one of the sections of mathematical linguistics – the theory of ways to describe syntactic structure.

Scientists consider mathematical linguistics primarily as a tool of theoretical linguistics. At the same time, its methods are widely used in applied linguistic research – automatic text processing, automatic translation and developments related to the so-called communication between humans and computers.

In this paper, we are interested in considering one aspect of mathematical linguistics, namely mathematical modeling of speech activity. The concept of a linguistic model originated in structural linguistics, but it became generally accepted in the 60-70s of the XX century with the emergence of mathematical linguistics and the penetration of ideas and methods of Cybernetics into linguistics.

Most scientists believe that a model in linguistics is a real or mental device artificially created by a linguist that reproduces or imitates (usually in a simplified form) the behavior of some other ("real") device (original) for linguistic purposes. Yu. D. Apresyan distinguishes three types of models, depending on the character of objects:

1) models of human speech activity that mimic specific language processes and phenomena;

2) models of linguistic research that mimic procedures that lead the linguist to the detection of one or another linguistic phenomenon;

3) the meta-model that simulates the theoretical and experimental estimation of ready-made models of speech activity or linguistic research.

Linguistic modeling necessarily involves the use of abstraction and idealization. Every model is built on the basis of the hypothesis about the possible structure of the original text and is its functional analogue. This approach allows you to transfer knowledge from the model to the original itself. A criterion for the adequacy of the model is a practical experiment.

Some researchers have noted that, any model should be formal (i.e., it should contain explicitly and unambiguously initial objects that relate their relationships andrules for dealing with them) and have explanatory power (i.e. not only explain facts or experimental data which are inexplicable from the point of view of an existing theory, but also to predict the previously unknown, although fundamentally possible behavior of the original, which should later be confirmed by observation or new experiments).

Depending on the aspect of language which is the subject of modeling, speech activity models are divided into *models of grammatical correctness*, imitating the ability to distinguish right from wrong in the language, and *functional,* imitating the ability to correlate the content of speech (content plan) with its form (expression plan). Depending on the type of information on the "input" and "output" models of grammatical correctness are divided into *recognizing* and *generating ones.*

*The recognizing model*receives at the «input» a certain section of text in natural language or its abstract representation in artificial language and gives an answer at the «output» whether this section is grammatically correct or abnormal.

The generating model is the reverse of the recognizing model. The generating model is the reverse of the recognizing model. Critical consideration of the first version of Chomsky's "generative grammar" led to the creation of a model of generative semantics which has much in common with models of speaking, or synthesis.

Depending on which aspect of speech activity is being modeled – listening or speaking-functional models are divided into *analytical* and *synthetic*. A complete analytical model of a certain language receives a certain section of text at the «input» (usually no less than a statement) and gives its semantic record (semantic

representation) at the «output» in a special semantic metalanguage (i.e., its interpretation). The complete synthetic model of a certain language, being the reverse of the full analytical model, receives a semantic record (an image of a certain fragment of meaning) at the «input», and at the «output» it gives a set of synonymous texts in this language that Express this meaning. Analysis and synthesis models form a necessary part of the translation model (in particular, the automatic translation model) and various "artificial intelligence" systems (in particular, question-and-answer systems).

The speech activity model is the most important type of linguistic models. In relation to them, models of linguistic research and metamodels play a supporting role. Research models are intended for objective justification of the choice of concepts that linguists use when presenting models of speech activity (for example, the grammar of a particular language). Ideally, they minimize the role of the subjective factor in the study and are in a sense a measure of the adequacy of models of speech activity.

Many Russian linguists in their works have investigated the problem of developing and applying mathematical models in various areas of linguistics. So, S.T. Starostin considered comparative historical linguistics and lexicostatistics, analyzed linguistic chronology and studied the time of the emergence of terms [6], S.V. Grinev-Grinevich using the mathematical model was able to determine the average time of origin of certain terms [2]. M.T. Dyachock analysed the language chronology and M. Swadesh, being the founder of the glottochronological method, was able to compile lexicostatistical lists that played important role in the development of linguistics in his work “Lexico-statistic dating of prehistoric ethnic contacts” [5]. At the same time, there is no doubt that studying the date of term origin and clarifying the language chronology requires additional research. The purpose of this article is to analyze the application of mathematical models in a particular area of linguistics – glottochronology. In the course of the study, both empirical and theoretical methods of scientific knowledge were applied: description and comparative analysis of using mathematical models in linguistics.

As it was mentioned earlier, there is currently an intersection of different areas of knowledge. According to researchesof domestic and foreign scientists, linguistics as a result of interaction with other branches of knowledge has become an extensive multidimensional science of comparative research of languages. Mathematics is one of the sciences closely intertwined with linguistics. It penetrates deeply into areas, which for a long time were considered to be purely “humanitarian”, expanding their heuristic potential. To find a connection between linguistics and mathematics, one of the areas of linguistics – glottochronology-should be considered.

Glottochronology, also called glottochronological lexicostatistics, is a research method. It was developed in the 1940s by the American linguist Morris Swadesh in order to date the time of divergence of related languages. The term "glottochronology" comes from the Greek *glossa* «language» and literally means «*language chronology»*. The term "glottochronology" is formed from the Greek. glossa 'language' and literally means 'language chronology' [7].

According to M. Swadesh, there is a list of words in the vocabulary of all the languages of the world that are distinguished by some stability. The linguist has developed a list of such terms: it includes the names of body parts, some living things, natural phenomena, some pronouns, adjectives that denote elementary geometric concepts and colors, a number of verbs, and some other words. Lists were made of 100 and 200 elements, called Swadesh's hundred-word and two-hundred-word lists [4]. Swadesh thought that the percentage of words from this list that will pesist for a certain period of time, which means that they will not be replaced with other words, is approximately the same for all languages: from a 200 word list, after 1000 years, approximately 81% of words are saved; from 100 word – about 86%.Nevertheless, the researchers argue that the lexicostatistical list needs to be specified and clarified, especially in cases where the English word allows different interpretations. This is necessary to avoid ambiguities and related errors. The work of clarifying the semantics of such words was started by M. Swadesh himself, and later continued by other researchers, in particular S.A. Starostin [6].

Considering glottochronology as a way to determine the age of divergence of languages by the number of words that have a common origin in them, let's consider the case of splitting the source language into two and try to determine the minimum time of divergence of these languages [5].

According to M.T. Swadesh the rate of saving or not saving words during a selected time is relatively constant, as is the speed of radioactive decay. For any period of time, the possible number of words, which after this period will be saved at the same in two languages, can be calculated, and vice versa, for any number of words that coincide in two related languages, the probable time elapsed after the divergence of the two languages can be calculated. The researcher determined the time with the formula:

$$d = \frac{log\ C}{2 * log\ r} \qquad (1)$$

where C is the part of the remaining words from the list, and r is the preservation coefficient equal, respectively, 0,81 or 0,85.

As an example the linguist presents the data on the approximate of divergence time depending on the percentage of the remaining vocabulary in languages, calculated by the formula (1).

APPROXIMATE TIME OF DIVERGENCE DEPENDING ON THE PERCENTAGE OF PRESERVED VOCABULARY IN LANGUAGES

Table 1.

№	Percentage of vocabulary preserved, %	Time of divergence, years
1	95	100
2	90	250
3	85	400
4	80	500
5	75	650
6	70	800
7	65	1000
8	60	1200
9	55	1450

10	50	1700
11	45	1900
12	40	2150
13	35	2500

Many researchers believe that the lexicostatistical lists proposed By M. Swadesh can be considered not accurate [4]. Discussions continue among critics of the glottochronological method about adding a particular word to the list, in a specific meaning. At the same time, critics admit that the problem can be solved, if we recognize the rather relative nature of the listand its reliability, proven during numerous attempts to apply the method.

Based on the analysis of the percentage of similarity of languages by the two-word list, the following table is compiled.

ESTIMATED TIME OF DIVERGENCE OF SOME LANGUAGES

Table 2.

Languages	The percentage of matching vocabulary, %	Estimated time of divergence of languages, years
RussianandUkrainian	89	290
ItalianandPortuguese	79	530
FrenchandSpanish	56	1390
EnglishandGerman	37	2350

With the help of the formula, possible time of language divergence was calculated. If we compare figures with the data in Table 1, we see that the calculated periods coincide with the periods in it.The result of mathematical analysis of the divergence of languages which have close cultural relations suggests and proves the undoubted importance and significance of strengthening inter-state relations, studying the culture and traditions of various countries and peoples, as well as the need to continue studying the deep structures of languages of the world using innovative methods of digital technologies.

Initially, mathematical methods in linguistics were used to clarify the basic concepts of linguistics, to determine the period of the emergence and subsequent

separation of languages of the peoples of the world. However, with the development of computer technology, such a theoretical premise began to find application in practice. The solution of such tasks as machine translation, searching information, automatic text processing requires a fundamentally new approach to the language. Therefore, the joint development of mathematics and linguistics can give great results. Linguistics is becoming more and more accurate and more objective science without ceasing to be a humanitarian science.

Model construction is not only one of the means of displaying language phenomena and processes, but also an objective practical criterion for verifying the truth of our knowledge of language. Applied in organic unity with other methods of language learning, modeling acts as a means of deepening the knowledge of hidden mechanisms of speech activity, its movement from relatively primitive models to more meaningful models that more fully reveal the essence of language.

Литература

1. ГЕО. Лингвистика. Список Свадеша. // URL: *https://geo.koltyrin.ru/spisok_svodesha.php* (дата обращения: 05.04.2019).

2. *Гринев-Гриневич С.В.* Терминоведение. Учебное пособие для студ высших учебных заведений, 2008 // URL: https://search.rsl.ru/ru/record/01004146309 (дата обращения 25.02.20).

3. *Дьячок М.Т.* Глотохронология пятьдесят лет спустя. Сибарский лингвистический семинар, 2002 // URL: http://philology.ru/linguistics1/dyachok-02b.htm (дата обращения: 16.04.2019).

4. *Звегинцев В.А.* Лингвистическое датирование методом глоттохронологии (лексостатистики). Новое в лингвистике. Вып. 1, М., 1960. S. 9-22. // URL: http://philology.ru/linguistics1/zvegintsev-60.htm (accessed: 16.04.2019).

5. *Свадеш М.* Лексикостатистическое датирование доисторических этнических контактов (на материале племен эскимосов и североамериканских

индейцев). Новое в лингвистике. Вып. 1. М., 1960. С. 23–52. // URL: http://philology.ru/linguistics1/swadesh-60.htm (дата обращения: 04.05.2019).

6. *Старостин С.А.* Сравнительно-историческое языкознание и лексикостатистика. Лингвистическая реконструкция и древняя история Востока, 1989 // URL: http://www.garshin.ru/linguistics/historical/comparative-books.html (дата обращения: 25.02.20)

7. Энциклопедия «Кругосвет». Глотохронология. // URL: https://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistik a/GLOTTOHRONOLOGIYA.html (датаобращения: 27.03.2019).

References

1. *GEO. Linguistica. Spisok Swadesha* [GEO. Linguistics. Swadeshlist]// URL:https://geo.koltyrin.ru/spisok_svodesha.php (accessed: 05.04.2019). (In Russian)

2. Grinev-Grinevich, S.V. (2008). *Terminovedenie.*Ucheb. posobiedlya stud. vysh. ucheb. zavedeniy/ Izdatelskiydom «Akademiya» [Terminology. Textbook for students of higher education institutions], 304 s. // URL: https://search.rsl.ru/ru/record/01004146309 (accessed: 25.02.20). (in Russian)

3. Dyachok, M.T. (2002). *Glotokhronologiya pyat'desyat let spustya. Sibirskii lingvisticheskii seminar.* [Glottochronology: fifty years later. Siberian linguistic seminar] // URL: http://philology.ru/linguistics1/dyachok-02b.htm (accessed: 16.04.2019). (In Russian)

4. Zvegintsev, V.A. (1960). *Lingvisticheskoe datirovanie metodom glottokhronologii (leksikostatistiki)* [Lexicostatistical dating by glottochronology (lexicostatistics)]. Novoe v lingvistike. Vyp. 1. M., 1960. S. 9-22. // URL: http://philology.ru/linguistics1/zvegintsev-60.htm (accessed: 16.04.2019). (In Russian)

5. Swadesh, M. (1960). *Leksikostatisticheskoe datirovanie doistoricheskikh etnicheskikh kontaktov (na material plemen eskimosov I severoamerikanskikh indeitsev*) [Lexicostatistic Dating of Prehistoric Ethnic Contacts with Special

Reference to North American Indians and Eskimos]. Novoe v lingvistike. Vyp. 1, M., 1960, S. 23–52. // URL: http://philology.ru/linguistics1/swadesh-60.htm (accessed: 04.05.2019). (In Russian)

6. Starostin, S.A. (1989). Sravnitel'no-istoricheskoe yazykoznanie i leksiko-statistika // Lingvisticheskaya rekonstruktsiya i drevneishaya istoriya Vostoka. [Comparative historical linguistics and lexicostatistics // Linguistic reconstruction and ancient history of the East. Materials for the discussions of the international conference] // URL: http://www.garshin.ru/linguistics/historical/comparative-books.html (accessed: 25.02.20). (In Russian)

7. *Entsiklopediya "Krugosvet". Glottokhronologiya.* [Encyclopedia «Krugosvet». Glottochronology] // URL: https://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/GLOTTOHRONOLOGIYA.html (accessed: 27.03.2019). (In Russian)

Авторы публикации

Сигачева Наталья Альбертовна – *кандидат педагогических наук, доцент*
Казанский федеральный университет
г. Казань, Россия.
E-mail: nsigacheva@mail.ru;

Баранова Альфия Рафаиловна *– кандидат педагогических наук, доцент*
Казанский федеральный университет
г. Казань, Россия
E-mail: baranova.alfiyarafailona@mail.ru

Authors of the publication

Sigacheva Natalya Albertovna *– candidate of pedagogical sciences, Associate Professor, Kazan Federal University*
Kazan, Russia
Email: nsigacheva@mail.ru;

Baranova Alfiya Rafailovna *– candidate of pedagogical sciences, Associate Professor, Kazan Federal University*
Kazan, Russia
Email: baranova.alfiyarafailona@mail.ru