ЛИНГВИСТИКА И СТАТИСТИКА

Н.А. Каменева, А.Ю. Широких

Статья посвящена специфике использования статистических методов в лингвистических исследованиях. Авторами отмечена важность адаптации статистических методов к решению филологических проблем и исследованию таких областей языкознания, как лексикография, культура речи, расшифровка древних текстов, установление авторства литературных произведений, методика преподавания языков и т.д. Сделан вывод, что использование методов статистического анализа в языкознании требует от лингвиста и переводчика владения не только знаниями языков и лингвистической проблематикой, но и до некоторой степени аппаратом математической статистики.

КЛЮЧЕВЫЕ СЛОВА: лингво-статистические методы, лингвостатистика, стилистика, квантитативная лингвистика, лингвистические закономерности

КАМЕНЕВА Наталия Александровна – кандидат экономических наук, доцент кафедры иностранных языков и профкоммуникации Московского финансово-юридического университета МФЮА. n-kameneva@yandex.ru; 29181420@s.mfua.ru

ШИРОКИХ Анна Юрьевна – кандидат филологических наук, доцент, доцент Департамента языковой подготовки Финансового университета при Правительстве РФ. ashirokih@mail.ru; ayshirokih@fa.ru

Цитирование: Каменева Н.А., Широких А.Ю. Лингвистика и статистика // Мир лингвистики и коммуникации: электронный научный журнал. 2018. № 1. С. 96–104. Режим доступа: www.tverlingua.ru

LINGUISTICS AND STATISTICS

Natalia A. Kameneva, Anna Yu. Shirokikh

The article is devoted to the specifics of using the statistical methods in linguistic studies. The authors stress the importance of statistical methods for the

solution of philological problems and research in such linguistic fields as lexicography, speech culture, deciphering ancient texts, attribution of literary works, methods of language teaching, etc. It is concluded that doing research in different linguistic disciplines stipulates the use of statistical analysis methods. It leads to a linguist's acquisition of new, mathematical skills.

KEY WORDS: lingvo-statistical methods, lingvostatistics, stylistics, quantitative linguistics, linguistic regularities

KAMENEVA Natalia A. – PhD in Economics, Associate Professor of the Department of foreign languages and professional communication of Moscow Financial and Law University MFUA. n-kameneva@yandex.ru; 29181420@s.mfua.ru

SHIROKIKH Anna Yu. – PhD in Philology, Associate Professor of the department of language studies of Financial University under the Government of the RF. ashirokih@mail.ru; ayshirokih@fa.ru

Citation: Kameneva N.A., Shirokikh A.Yu. Linguistics and Statistics // World of linguistics and communication: electronic scientific journal. 2018. № 1. P. 96–104. Access mode: www.tverlingua.ru

Statistical methods in linguistics represent a set of techniques and principles designed to facilitate the process of collection, classification and interpretation of verbal data in order to obtain scientific and practical conclusions. These conclusions may be applicable in discussing how the language develops and what speech structures are more or less frequently used. Thus, linguists, interpreters, philologists, literary scholars, teachers, and psychologists take special interest in mathematical statistics tools.

The statistical linguistic data is quantitative information about a particular set of language phenomena that share common properties. Conventionally, language statistics methods fall into two groups which may be further detailed as primary processing methods and secondary processing methods:

1. Descriptive methods, i.e. those which portray a variety of texts and present the obtained data in a compact and understandable form (e.g. in the form of tables or graphs). These methods use different kinds of statistics (e.g. type of communication indicators, frequency indicators, measures of central tendency)

2. Methods of evaluation are studied in the framework of such sciences as comparative linguistics, typological linguistics, and quantitative linguistics.

The statistical method of analysis is considered a universal method of cognition, since in comparison to other methods of research it has such undeniable advantages as objectivity, impartiality, rigor and procedurality. Therefore, the statistical analysis method is actively and universally used in different sciences for compact representation, analysis, synthesis and interpretation of the empirical data. Such linguistic categories of "language", "speech", "text" can be reinterpreted through statistical representation models and “old” linguistic problems may be seen in the new light. Hence, not only are linguists required to be experts in communication process but also possess some mathematical skills and knowledge of statistics methodology.

Speaking about the role of statistics in language and style description patterns, we are to remember the achievements of the famous Russian linguist, Honored Scientist of the USSR, Doctor of Philological Sciences, Professor Boris Nickolayevich Golovin. B. N. Golovin is reasonably considered one of the founders of modern lingvostatistics. Text analysis being a sphere of his scientific interest, he was among the first who drew attention to the practical importance of statistical methods for examining linguistic regularities. His research contributed much to the theory of languages.

Golovin stated that language functioning in the speech is subject to the statistical laws and statistical nature, i.e. our speech is formed in accordance with certain statistical laws (Golovin, 1970: 16-17).

Functional stylistics considers utility of verbal structures for different types of discourse. Communication modes are influenced by numerous linguistic and

extra-linguistic factors. To quantify verbal performance statistical methods can be used. Statistics is also applicable for research in:

– lexicography, concerned with compilation of frequency dictionaries, national corpora etc.;

– speech culture, that defines language norms;

– prosody, that is the study of poetic rhythm and the art of versification;

– interpretation of ancient texts;

– teaching methods (compilation of minimum vocabulary lists for students, measurements of classroom performance in different educational environment, development of curricula etc.) ;

– attribution of literary and scientific texts and consequently indexing a scholar's works, etc.

There is an opinion that science is perfected only when using precise mathematical methods. The theoretical foundation of the quantitative analysis methods and the creation of algorithms for their practical application in linguistics is the subject of lingvostatistics as a special branch of science.

The essence of the lingvo-statistical method is to establish the quantitative changes that cause the qualitative transformations in the linguistic phenomena. Through applying mathematical methods for investigations in stylistics it was discovered that the frequency of linguistic elements is subject to certain laws or regularities. It leads to a conclusion that language is a living organism and a living organism is contingent on its behavior in different environments. Speech in action is discourse and discourse can be measured and analyzed through statistical representations. Recently, a large number of electronic and on-line dictionaries including different analyses of modern linguistic phenomena have appeared (Kameneva, 2015: 87).

For example, when using the lingvo-statistical method the experts study the quantitative characteristics of the vocabulary of the different speech styles, the discourse types and the varieties of the author's speech. As a result frequency

dictionaries are compiled and published. An example can be the Frequency Dictionary of Modern Russian Language by Lyashevskaya O.N., Sharov S.A. Their dictionary contains variable statistical data for 50 000 words of general vocabulary, inherent to journalism, fiction and other functional linguistic styles (3). This dictionary is based on the collection of texts of the National Corpus of the Russian language, representing the modern Russian language for the period 1950-2007. The sample size on which most sections of the dictionary are built is 92 million words (3). The frequency list of lemmas in this dictionary provides information on word frequency for different decades of the second half of the XX century and the beginning of the XXI century. For example, we can see the frequency rate for the use of the Russian word "perestroika" (see Table) (Lyashevskaya, Sharov, 2009 online resource).

Table

Period	1950s	1960s	1970s	1980s	1990s	2000s
The frequency of usage for "perestroika" per a million samples	7,51	3,89	8,42	48,50	23,17	31,29

The increase in the frequency of the word "perestroika" in the 1980s is caused by social and historical trends of that time. At the same time, from linguistic point of view, this fact can be interpreted as follows: the word "perestroika" got a new meaning, which has become predominant in the coming years: "The policy of reconstructing the economy, etc., of the former Soviet Union under the leadership of Mikhail Gorbachov" (Online Dictionary ABBYY Lingvo, 2018).

The distribution of lexical items frequency is extremely uneven: there are some rare words that are used more than common every day vocabulary. Zipf's Law defines the inversely proportional relationship between the sequence number in the frequency word list (r, rank) and frequency (f):

$f \approx kr^{-\alpha}$, where k - constant depending on the case (absolute number of uses of the frequency of the word),

and α - close to one, a mathematical power parameter, depending on the grammatical structure of the language.

Quantitative description of sublanguages is used in certain thematic areas. It allows for automatic processing of linguistic information in development of information retrieval systems or methodology of language teaching. It is implemental for identifying thematic vocabulary for students' memorization.

Zipf's law is simply an attempt of applying the Pareto Principle to human languages, and we can estimate its applicability. Data suggests that 20% of words of a language (such as English, Russian, German, Spanish, Mandarin or Belarusian) substitute 80% of a speaker's active vocabulary, i.e. what this person ever says or writes (BlogEMIS, 2018 online resource).

Nowadays a detailed analysis of a language is inconceivable without modern rigorous quantitative estimates. Well-known scientists such as I.A. Baudouin de Courtenay, A.M. Peshkovski, M.N. Peterson, E.D. Polivanov, V.V. Vinogradov and others devoted their scientific works to the quantitative estimates of the language elements.

As early as 1938, V. Vinogradov wrote that in different bookish styles and oral speech, as well as in different genres of literature the frequency of vocabulary is diverse (Golovin, 1970: 11).

The accurate research in this area would help to establish the structural grammatical and semantic differences between speech styles and discourses. However, unfortunately, sometimes this issue is only a preparatory stage for examining texts. Any modern live language undergoes continuous changes. The vocabulary once considered colloquial may become quite bookish in style a century later. The same is true for grammar structures. History of language development indicates shifts in tense and voice structures of the English language, for instance. Analysis of grammatical categories helps us understand their relative

functional weight in different styles of the literary language (Khokhlova, 2015: 176).

Profound scientific study of language and speech is impossible without comprehensive array of statistical data. The styles are distinguished because each of them serves a different aspect of our life. These language styles are singled out based on their probabilistic characteristics of the same linguistic element in a certain language structure.

Several statistical elements, such as the frequency of words, phonetic patterns, sentence structures, word formation, etc. can provide the basic information about the structure of the language to explicit the message of a text.

Students can use the statistical data to determine the most probable meaning of the lexical unit and limitations these meanings have in a particular context. The inductive learning method may be practiced if we start with studying the contexts and then try to make conclusions about the rules of this word's functioning. Ambiguous situations, figurative meanings, possible references or allusions can lead to a student's confusion. This is statistics and its products, e.g. dictionaries or web sites, which help the students resolve such problems. The same is true for phonetics and grammar structures – statistics systemizes knowledge and represents it in an accessible way.

Modern linguists have become accustomed to think in terms of quantities, probabilities and trends in determination of such speech properties as length, frequency, time of occurrence, the degree of ambiguity of an utterance. These are universal properties studied by statistics.

Ссылки – References in Russian

BlogEMIS online resource, 2018 – *BlogEMIS* // https://blogemis.com/2015/09/26/zipfs-law-and-the-math-of-reason/

Golovin, 1970: 16-17 – *Головин Б.Н.* Язык и статистика. М.: Просвещение, 1970.

Kameneva, 2015: 87 – *Каменева Н.А*. Компьютерная лексикография и составление электронных словарей. Филологические науки. Вопросы теории и практики. 2015. № 3-1(45). С. 86–89.

Khokhlova, 2015: 176 – *Хохлова М.В.* Большие корпусы и частотные существительные: предварительные наблюдения. В сборнике Структурная и прикладная лингвистика. Межвуз. сб. / Под ред. А. С. Герда и И. С. Николаева. Вып. 11. СПб.: Изд-во С.-Петерб. ун-та, 2015. С. 174–185.

Lyashevskaya, Sharov, 2009 online resource – *Ляшевская О.Н., Шаров С.А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009 // http://dict.ruslang.ru/freq.php?act=show&dic=freq_freq

Online Dictionary ABBYY Lingvo Live, 2018 – Онлайн-словарь ABBYY Lingvo (электронный ресурс) Режим доступа: https://www.lingvolive.com/ru-ru (Дата обращения 5.01.2018).

References

BlogEMIS, (2018) // https://blogemis.com/2015/09/26/zipfs-law-and-the-math-of-reason/

Golovin, B. (1970) *Language and Statistics*, M., Prosveshcheniye (in Russian).

Kameneva, N. (2015) *Computer lexicography and compilation of electronic dictionaries*, Philological Sciences, Issues of theory and practice, № 3-1(45), pp. 86–89. (in Russian).

Khokhlova, M. (2015) *Large cases and frequency nouns: preliminary observations.* In the collection of structural and applied linguistics, Issue. 11:

interuniversity collection / Ed. AS Gerda and IS Nikolaev, St. Petersburg, Publishing house S.-Petersburg, University, pp. 174–185. (in Russian).

Lyashevskaya, O., Sharov, S. (2009) *Frequency Dictionary of Modern Russian Language (on materials of the Russian National Corpus)*, M., Azbukovnik // http://dict.ruslang.ru/freq.php?act=show&dic=freq_freq (in Russian).

Online Dictionary ABBYY Lingvo (2018) // https://www.lingvolive.com/ru-ru