UDC 81‘32

# STATISTICAL METHODS OF FORMATION OF TEXT CORPORA AND LEXICOGRAPHIC RESOURCES (ON THE BASIS OF THE SPECIALTY “ACOUSTICS AND ULTRASONIC”)

# СТАТИСТИЧНІ МЕТОДИ ФОРМУВАННЯ ТЕКСТОВИХ КОРПУСІВ І ЛЕКСИКОГРАФІЧНИХ РЕСУРСІВ (НА ПРИКЛАДІ СПЕЦІАЛЬНОСТІ «АКУСТИКА І УЛЬТРАЗВУКОВА ТЕХНІКА»)

**Dyachenko G.F.,**
*PhD, Associate Professor,*
*Associate Professor of the Foreign Language Department*
*Odessa National Polytechnic University*

**Mykhailiuk S.L.,**
*Senior Lecturer of the Foreign Language Department,*
*Odessa National Polytechnic University*

**Duvanskaya I.F.,**
*Senior Lecturer of the Foreign Language Department,*
*Odessa National Polytechnic University*

**Ershova Yu.A.,**
*Senior Lecturer of the Foreign Language Department*
*Odessa National Polytechnic University*

The article considers the description of the step sequence in forming the text corpora, and then frequency dictionaries on the example of Acoustics and Ultrasonic Technique (AUST) specialty, the texts of which are referred to scientific and technical discourse. The necessity of application of real text corpora compiled with the help of statistical methods in the present-day research processes is proved. Statistical method usage allows to determine such a mandatory parameter as the reliability of text corpus and lexicographic resources created on its basis – frequency dictionaries, alphabet-frequency dictionaries, etc. The example of specialty AUST demonstrates how statistically verified characteristics of the text corpus allowed to create a reliable probabilistic-statistical model (frequency dictionary) of this subject area. The statistical reliability of the dictionary manifested itself in the fact that the percentage of covering the AUST texts with the units of the base dictionary (the first 2 thousand words) is 86%, which makes it possible to understand the content of almost any text on the specialty AUST using the lexical units presented in it (the base dictionary).

**Key words:** absolute frequency, relative frequency, probabilistic-statistical model, subtopic, text corpus.

Стаття присвячена опису послідовності кроків під час формування текстових корпусів, а потім – частотних словників на прикладі спеціальності «Акустика і ультразвукова техніка» (АУЗТ), тексти якої належать до науково-технічного дискурсу. Доводиться необхідність застосування в сучасних дослідних процесах реальних текстових корпусів, створених на основі статистичних методів. Використання статистичних методів дозволяє визначити такий обов’язковий параметр, як надійність текстового корпусу й створюваних на його основі лексикографічних ресурсів – частотних словників, алфавітно-частотних словників та ін. На прикладі спеціальності АУЗТ було показано, як статистично вивірені характеристики текстового корпусу дозволили створити надійну ймовірносно-статистичну модель (частотний словник) цієї предметної області. Статистична надійність словника проявила себе в тому факті, що частка покриття текстів АУЗТ одиницями базового словника (перших 2 тис. слів) склала 86%, що дає можливість за допомогою лексичних одиниць, представлених у базовому словнику, зрозуміти зміст практично будь-якого тексту за фахом АУЗТ.

**Ключові слова:** абсолютна частота, відносна частота, ймовірносно-статистична модель, підтема, текстовий корпус.

Статья посвящена описанию последовательности шагов при формировании текстовых корпусов и частотных словарей на примере специальности «Акустика и ультразвуковая техника» (АУЗТ), тексты которой входят в научно-технический дискурс. Доказывается необходимость применения в современных исследовательских процессах реальных текстовых корпусов, созданных на основе статистических методов. Использование статистических методов позволяет определить такой обязательный параметр, как надежность текстового корпуса и создаваемых на его основе лексикографических ресурсов – частотных словарей, алфавитно-частотных словарей и др. На примере специальности АУЗТ было показано, как статистически выверенные характеристики текстового корпуса позволили создать надежную вероятностно-статистическую модель (частотный словарь) этой предметной области. Статистическая надежность словаря проявила себя в том факте, что доля покрываемости текстов АУЗТ единицами базового словаря (первых 2 тыс. слов) составила 86%, что дает возможность с помощью лексических единиц, представленных в базовом словаре, понять содержание практически любого текста по специальности АУЗТ.

**Ключевые слова:** абсолютная частота, вероятностно-статистическая модель, относительная частота, подтема, текстовый корпус.

The main tendencies in modern linguistics can be considered the following: computer, corpus and cognitive linguistics. All of them provide for the researches on the material of real texts, i.e. the text corpus sorted out from all set of texts on the chosen subject – technical, humanitarian, art, etc. – with possible further compilation of probabilistic and statistical model (the frequency dictionary or frequency list) of that area of a discourse which was chosen for the analysis.

However, it is believed that not all of the researchers know how to apply quantitative and qualitative methods when forming both the semantic space, which is necessary for modeling the semantics of a certain area of knowledge, and the future frequency dictionary.

This article based on the example of the frequency dictionary of the technical specialty "Acoustics and Ultrasonic Technique" created by the authors describes the sequence of steps which are carried out to receive statistically reliable objects of the research.

Works on the basis of text corpora became especially widespread in the XXI century, and the most various text units analyzed both in respect of a lexicography, and corpus linguistics: terminology [1; 2; 3; 4; 5], stratification layers of specialties [6; 7], principles of the statistical analysis [8; 9], etc. served as objects for consideration.

Despite a significant amount of the works devoted to the analysis of text objects, it is possible to note that, unfortunately, they do not always contain statistically verified information on text corpora, methods of their formation and further use. So, the researcher Ivina [3], representing a subject of the scientific work, mentions only those concrete speech units which were taken from the considered texts, but not their statistical data. Or, despite a significant amount of the works devoted to the analysis of text objects, it is possible to note that, unfortunately, they do not always contain statistically verified information on text corpora, methods of their formation and further use. So, the researcher Ivina [3], representing a subject of the scientific work, mentions only those concrete speech units which were taken from the considered texts, but not their statistical data. Or, though the description of speech units is declared in the majority of scientific research, they do not rely on the text corpus created by the method of continuous sampling, but on lexicographic sources torn off from the real scientific speech, in which it is possible to observe the various elements of the speech only [1; 2; 10].

Due to the above said, i its believed that the description of the corpus creation process of the chosen specialty "Acoustics and Ultrasonic Technique", and then – probabilistic-statistical model of this specialty on the basis of both theoretical sources [11; 12; 13], and experimental simulation, is relevant and timely, and will be a useful example for researchers in the field of corpus linguistics and a statistical lexicography.

So the goal of the given article is to describe in detail the statistical methods and practical steps in the course of the text corpus compilation of the "Acoustics and Ultrasonic Technique" specialty and as well as the further formation of its probabilistic-statistical model.

The first task in distinguishing the text corpus is to define the so-called semantic space of the chosen subject area, in our case – "Acoustics and Ultrasonic Technique" (AUSE) which models its (area) semantics.

It is clear that to statistically survey all texts of the AUSE sublanguage does not seem possible. Therefore, we were guided by some reduced description of sublanguage, i.e. its linguistic model constructed as some statistical approximation to the real lexical and grammatical system of the AUSE specialty.

To create the semantic space of fields of knowledge which belong to scientific (humanitarian or technical) discourse, is frequently a rather difficult procedure as researchers-linguists cannot be absolutely professionally competent in many technical specialties. In this case, they should rely on the opinions of experts and resort to a method of expert assessment as well as use other sources of description.

In forming semantic space of the given area it has been revealed that it breaks up to six subareas (subthemes) which were further stemmed as a result of journal articles inspection, studying references and poll of the experts who are engaged in studying of various aspects of acoustics and the ultrasonic technique. During the investigating of journal articles a need to present all sub-themes of this area, a per cent of word tokens in this or that sub-theme in a total amount of text set as well as the degree of importance of a subtheme within "Acoustics and Ultrasonic Technique" science was considered. Sub-themes are presented in a percentage correlation which they occupy in a total amount of AUSE texts. They are five:

– Ultrasonic instruments and devices – 30%;
– Hydro acoustics – 20%;
– Acoustic and ultrasonic signals and their processing – 20%;
– Acoustic and ultrasonic measurements – 15%;
– Designing, and acoustic and ultrasonic equipment technology – 10%;
– Noise and vibrations – 5%.

According to the content, nomenclature and fractions which sub-themes occupy in semantic space, first, the corresponding scientific journals, in which the above-mentioned sub-themes are described, were sorted out, second, the number of word tokens were

strictly calculated to corresponded to each fraction in the general list.

Figures, symbols, advertisement, popular scientific articles were not included into the text corpus. To reflect the current stage of development of AUSE, the corpus was constructed on chronologically limited material, i.e. it includes the texts limited to a 10-year interval. Texts were taken from the magazines issued in the USA and the UK: Journal of Acoustic Society of America, Journal of the Audio Engineering Society, Applied Acoustics, IEEE Transactions of Antennae and Propagation, The Journal of the Society of America, etc.

Selection of texts is of very great importance in compiling a frequency dictionary since the studied corpus represents that model which reflects statistical features of the sublanguage in the aspect of its lexical and grammatical structure.

Although there are several approaches to text corpus formation the complete scientific and technical articles, irrespective of the text length in word tokens, were used by the authors for the research only. It gave the chance to receive more reliable data on the material studied. Of course, our method of text selection is not deprived of some shortcomings. In particular, in high- and mid-frequencies areas of the future frequency list, the lexical units, which are not of big importance for this specialty, can be registered. However, these facts do not distort statistical structure of the text.

Assessment of reliability of the future frequency dictionary depends both on the qualitative, and quantitative characteristics which are considered when forming the text corpus [11; 14]. The quantitative characteristics provide for certain methods of text selection, qualitative – the content of material of the studied text set, i.e. their strict reference to the distinguished sub-themes in the semantic space.

The quantitative description of reliability of the frequency dictionary also depends on the volume of text corpus that complies with the earlier set criteria of reliability of a future frequency dictionary. The volume of our corpus is 200 thousand word tokens. In the course of creating the frequency dictionary of high percentage of covering the text corpus contents, it must be kept in mind that the volume of corpus depends on a language system and the discourse a text is referred to. For analytical languages, by compiling branch frequency dictionaries on the basis of scientific and technical texts of the certain subjects, the corpus volume in 200 thousand word tokens is considered to be traditional and sufficient for covering the branch literature texts for 97–99% [11; 14; 15].

The following stage is, actually, the use of the created text corpus for various researches including those where statistics are necessary.

As the modern linguistics requires application of not only linguistic (contextual, distributive, grammatical, etc.) analysis methods, but also the exact indication of quantitative indices at representation of these or those facts of language, the mathematical approaches included in linguistic research is its obligatory component.

First of all, in order to obtain the quantitative data, it is necessary to create the frequency list (or if to apply the term accepted in linguistic statistics – probabilistic-statistical model of specialty) of all word forms or word tokens which occur in texts. It is the simplest probabilistic-statistical model of AUSE sublanguage lexis.

There are programs which arrange all word tokens according to their frequencies (on frequency decrease) or in an alphabetic order. The frequency AUSE dictionary was created in the same way .

The frequency dictionary possesses the following statistical characteristics: absolute frequency of the use of this or that word token (F), relative frequency (f) which is calculated by the simple ratio of length of the corpus to the frequency of a word token and which allows the researcher to use data of the dictionary for comparison with the data received, for example, at analysis of the text corpus bigger (or smaller) than 200 thousand word tokens. Besides, at each unit of the dictionary, an absolute cumulative frequency $F^*$ is specified as the sum of the presented frequency $F$ and all frequencies preceding it and also the cumulative relative $f^*$ frequency equal to the sum of this $f$ and all previous relative frequencies.

The total amount of the frequency dictionary is up to 5 648 different words. The experience of dictionary compilers demonstrates that a dictionary comprising approximately 2,000 lexical units is sufficient to cover the contents of branch literature texts for 89% [16]. Conventionally, such part of the dictionary is called the base dictionary. In our case, the word tokens of the base dictionary cover all AUSE lexicon up to 86%. It shows that the base frequency list contains the lexical units minimum, which is enough for decoding the texts on acoustics and the ultrasonic technique, and a reader has an opportunity to completely understand the contents of the text in AUSE. In this case, it is possible to say that the frequency dictionary compiled as a result of the statistical analysis, represents certain probabilistic-statistical model of specialty “Acoustics and the Ultrasonic Technique”.

The procedure of text corpus formation and then on its base – probabilistic-statistical model of specialty “Acoustics and the Ultrasonic Technique” described in the paper allows to draw the following conclusions.

The procedure of text corpus formation and then on its base – probabilistic-statistical model of specialty "Acoustics and the Ultrasonic Technique" described in the paper allows to draw the following conclusions.

1. Consideration of this or that speech unit is possible only on the basis of real text corpus.

2. To create the text corpus, it is necessary to form a so-called semantic space of specialty. For AUSE specialty, the semantic space which contains six sub-themes and which represents the main problems studied by this area of expertise was created.

3. According to a statement that the text corpus has to be compiled with obligatory application of the quantitative and qualitative methods which provide the statistical reliability of both the corpus and future probabilistic-statistical model of a certain specialty, for text corpus creation of AUSE specialty the following methods were used: the texts which entered the corpus cover the ten-year period; texts for the corpus were selected strictly according to fractions which sub-themes occupy in a total amount of word tokens; the volume of the corpus was 200 thousand word tokens that is sufficient for statistical reliability of linguistic objects.

4. Statistical characteristics of the AUSE frequency dictionary are the following. It contains 5,648 units which had the statistical parameters: absolute frequency of usage, relative frequency of usage; cumulative absolute and relative frequencies.

5. The value of covering the AUSE texts with the basic dictionary units (the first 2,000 words) is up to 86% that gives the chance to virtually understand the contents of any texts in AUSE.

In future, it is supposed to compile perhaps bigger number of text corpora that are various on their scientific subjects in order to compare both statistical and lexical data.

**REFERENCES:**

1. Sytnikova T. English-speaking computer technical term system as object of a linguo-cognitive research: Author's abstract, thesis for PhD degree: specialty 10.02.04 German languages. Vladivostok, 2010. 22 p.

2. Chistyukhina S. Interindustry polysemanticism in the terminological system of modern English: Author's abstract, thesis for PhD degree: specialty 10.02.04 German languages. M., 2011. 22 p.

3. Ivina L. Nominative and cognitive research of an English-speaking term system of venture financing: Authors abstract, thesis for PhD degree: specialty 10.02.04 German languages. M., 2001. 22 p.

4. Oganesyan M. The comparative and translation analysis of the English and Russian medical terminology on genetics: Author's abstract, thesis for PhD degree: German languages 10.02.20 . M., 2003. 22 p.

5. Trifonova E. Polysemanticism of bank terms in English: Author's abstract: thesis for PhD degree: specialty 10.02.04 German languages / E. N. Trifonova. Omsk, 2004. 22 p.

6. Chupilina E. System properties of general scientific lexicon. System description of lexicon of the German languages. L.: GLU, 1985. P. 109–113.

7. Dyachenko G., Tsinovaya M., Sirotenko T. The verbs of common lexical layer in the texts of scientific style "Acoustics and ultrasonics". Odesky lingvistichny visnik, ONJA. Vol. 1. No. 9. 2017. P. 60–65.

8. Dyachenko G., Duvanskaya F., Mykhailiuk S. Principles of lexical stratification of vocabulary (on the material of verbal lexis of texts "Acoustics and Ultrasonics"). XIII International Scientific and Practical Conference "Modern Scientific Potential – 2017". Vol. 6 "Philological sciences" (Sheffield, Yorkshire, England). P. 42–47.

9. Nevreva M., Lebedeva E., Gvozd E. The statistics of nominative word formation in text corpora of scientific functional style. European Journal of Literature and Linguistics. Vienna: "East West" Association for Advanced Studies and Higher Education GmbH, 2016. No. 4. P. 31–34. (Austria).

10. Milyaeva L. A structural-semantic research of word-formation options of nouns in modern English (paradigmatic aspect): Author's abstract, thesis for PhD degree: specialty 10.02.04 The German languages. Pyatigorsk, 1984. 25 p.

11. Piotrovsky Rajmund G. Quantitative Linguistics. An International Handbook. Walter de Gruyter: Berlin. New-York. 2005. 1027 p.

12. Summers Della Corpus Lexicography – The importance of representativeness in relation to frequency / Della Summers. URL: www.pearsonlongman.com/.../pdfs/corpus-lexicography.pdf.

13. Krishnamurthy R. Corpus Lexicography. Birmingham: Aston University. Elsevier Encyclopedia of Language and Linguistics – 2nd Edition. URL: https://www.researchgate.net/publication/291110989_Corpus_Lexicography.

14. Alekseev P. Statistical lexicography (typology, compiling and application of frequency dictionaries). L.: Leningr. state. pedagogical. institute after A. Herzen, 1975. 120 p.

15. Bektayev K., Lukyanenkov K. About zones of distribution of units of a written language. Statistics of the speech and the automatic analysis of the text. L.: Science, 1971. P. 47–112.

16. Beresnev S. A research of lexicon of the German scientific and technical texts from a position of the recipient of the speech: Author's abstract, thesis of a Doctor's degree. L.: Academy of Sciences of the USSR. Institute of linguistics. Leningrad dpt, 1974. 35 p.