UDC 81'373.611
DOI https://doi.org/10.32782/2409-1154.2024.66.26

Nevreva M. N.,
Candidate of Philological Sciences, PhD,
Associate Professor at Foreign Languages Department
National University "Odesa Polytechnic"

Duvanska I. F.,
Senior Lecturer at Foreign Languages Department
National University "Odesa Polytechnic"

Mikeshova G. P.,
Senior Lecturer at Foreign Languages Department
National University "Odesa Polytechnic"

SEQUENCE OF PROCEDURES FOR OBTAINING THE RELIABLE RESULTS IN DETERMINING THE STATISTICAL CHARACTERISTICS OF LINGUISTIC OBJECTS

Summary. The paper presents the possible order of analyzing the statistical characteristics of a linguistic unit functioning in text corpus. As an object, a statistical characteristic of productivity of creating the new words with the help of derivative elements adding to noun rooting morphemes has been considered. The materials for analysis are the text corpora of three engineering specialties referring to scientific and technical discourse – "Chemical engineering", "Automotive engineering" and "Electrical Engineering". They were compiled on the basis of scientific article published in the journals of corresponding fields: Chemical Engineering Progress; Process Engineering; Machinery; Chemical and Process Engineering; Automobile Engineering, Auto Industry; IEEE Transactions on Power Apparatus and Systems, Proceedings of the Institution of Electrical Engineers. The sizes of "Chemical engineering" and "Electrical engineering" text corpora are 200 thousand tokens, the one of "Automotive engineering" is 300 thousand tokens. All the three text corpora fields are of different specialties which are not connected to each other with their topics. That gives the opportunity to get the generalized results that make it possible to take some integrating or differentiating characteristics as a stylistic marker. The article describes in detail the sequence of necessary steps to obtain the reliable data of the productivity of noun root morphemes. The methods applied in the process of researching are as follows: contextual analysis with the help of which all the derivatives were extracted from the texts; statistical methods of calculation, which were used in the course of creating the probabilistic-statistical models (frequency dictionaries), conversion of absolute values into relative ones; method of surveying the specialists experienced in the specialties taken as the material for compilation of the text corpora, etc. The main subjects of the research are the noun root morphemes occurring in the texts with the highest magnitude of frequency to make the description of the subject of analysis as complete and obvious as possible.

Key words: frequency, productivity, value, absolute and relative digits, text corpus.

Statement of problem. Literature review. The description of the results of research implies a certain sequence that the author chooses to present the data obtained in order not only to record new knowledge that characterizes the linguistic object but also, perhaps, to describe the original methods that help in research and the order of their application.

It is traditional now for the authors of studies, who are already well acquainted with the theoretical and applied issues of corpus linguistics, to use the methods of contextual analysis in their descriptions [1; 2; 3]. However, this does not exclude the analysis of dictionary data, i.e. language systems [4]. Since at the current level of development of linguistics, scientists have come to the conclusion that the study of text corpora separately from the language system embodied in dictionaries is not productive, many researchers have attempted to combine both types of sources, in which they (authors) can determine both the semantic structure of a word implemented in the text and possible changes in dictionary definitions influenced by the use of new meanings [5; 6]. If the object of research is a linguistic phenomenon functioning in text corpora of technical specialties that belong to the scientific and technical type of discourse, then quite often experts specializing in any field of technology are involved in the research [7].

We can note a number of works devoted to the description of statistical parameters, as well as the compilation of probabilistic-statistical models of technical specialties [8; 9; 10]. However, none of these works describes in detail the necessary and sequential stages of analyzing the statistical characteristics of any text unit.

Therefore, this work contains a feature of novelty.

Goal of the article. The article has the following goal – to describe an example of obtaining the reliable results if any statistical characteristic is considered. In this article, the object of description is the productivity possessed by the analyzed linguistic unit. In our case, we took for consideration such a characteristic as the productivity of root morphemes, on the basis of which nouns were formed.

The study involves text corpora of technical specialties referred to scientific and technical discourse, in which the nouns derived from root morphemes have been found (nouns having only a root morpheme are not analyzed, since they could not express derivativeness).

In order for the results to have a sufficient degree of reliability, as well as to determine the integral and differential parameters of the research object, it was decided to use the text corpus of not one specific specialty but three.

Thus the material for the article are the text corpora of three different specialties as for their subjects – "Chemical Engineering", "Automotive Engineering" and "Electrical Engineering". All texts were taken from articles in scientific journals of relevant specialties: "Chemical Engineering" from Chemical Engineering Progress; Process Engineering; Machinery; Chemical and Process Engineering; for the text corpus of the specialty "Automotive Engineering" the journals Automobile Engineering, Auto Industry; for the "Electrical Engineering" corpus, IEEE Transactions on Power Apparatus and Systems, Proceedings of the Institution of Electrical Engineers were reviewed. The text corpora "Chemical Engineering" and "Electrical Engineering" consist of 200 thousand tokens, the "Automotive Engineering" corpus includes 300 thousand tokens.

In addition to obtaining more reliable analysis results, as stated in the description of the purpose of the article, the method of using not one, but several text corpora in specialties that do not have common scientific and technical topics, suggests that such a choice can contribute to obtaining the generalized results that will make the integral or differential characteristics of root morphemes the distinctive markers for texts of scientific and technical discourse.

Based on the text corpora, three probabilistic statistical models (frequency dictionaries) of the mentioned specialties were compiled, in which all units were arranged in descending order of frequency of use. All of them were conditionally divided into frequency zones – high frequencies from the beginning of the list to frequency 7, and low frequencies from 7 to 1.

A survey of texts on the automotive industry allowed us to identify 6936 nouns for analysis; from the chemical engineering text corpus 6589 nouns; from texts on electrical engineering 5700 nouns.

Base material. We should note that root morphemes, due to their nature, required a completely different type of processing and systematization than suffixal or prefixal morphemes already described by the authors in other scientific articles [11]. If suffixes and prefixes can simply be listed with a mention of the frequency of use and at the same time be illustrated with the help of corresponding examples proving their statistical and lexical characteristics, then for root morphemes located in the center of the word and connected with almost all morphological units (suffixes and prefixes), the description has become significantly more complicated.

Another difficulty is that data from frequency dictionaries make it possible to represent only the frequency of use of a particular lexeme, its position and environment in the frequency list. To determine productivity much more complex procedures are required and, above all, a thorough study of context because word productivity is the ability to form the new words, which in our case will help determine the number of them (new words) formed by adding suffixes and prefixes, as well as other linguistic elements, to the root morpheme.

The procedure for determining productivity contains the following necessary steps.

The first step is to analyze text corpora to identify the entire inventory of root morphemes. In the three corpora, it was found that a total of 4520 different root morphemes occur in the nouns of all text corpora. Most of them (2576 units) are implemented in lexemes of root nouns, while another, smaller part (1944 units) is used in lexemes of derived nouns. Since the work is not aimed at studying root words, only the roots of derived nouns should be analyzed.

The second step is the procedure for entering each root morpheme into a separate file, where its morphological environment is recorded. The morphological environment, as mentioned above, is determined by the affixes with which the root morpheme is connected, as well as by other root morphemes in the formation of complex words. First, the absolute frequency of use of each created lexeme is recorded, then the total frequency of all the lexemes with a given root morpheme is calculated. The total number of formed lexemes allows us to determine in the future the productivity of the morpheme.

Let us give an example of derived units from one of the most frequent root morphemes 'mix'. The example presents not only general quantitative values, but also the number of lexemes located in different frequency ranges – with high frequency of occurrence (first digit in brackets) and low frequency (second digit in brackets).

1) mix-er, частота F = 131
2) mix-ing F = 108
3) mix-ture F = 96
4) mix-er-settler F = 27
5) back-mix-ing F = 10
6) pre-mix-er F = 3
7) pro-mix-ing F = 2
8) forward-mix-ing F = 1
9) micro-mix-ing F = 1
10) mix-idness F = 1

Thus the total number of words created is 10 (5; 5), and the repetition rate (frequency) is 280 (272; 8).

As can be seen from the example, the root morpheme 'mix' is combined with three prefixes (pre-, -pro-, micro-), with four suffixes (-er, -ing, -ture, -ness) and is a part of polylexemic words (mixer -settler, back-mixing, forward-mixing). It is quite clear that not all words of the substantive nest recorded in the file are formed directly from the root morpheme. But the meaning of the root morpheme is "core", determining the unity of these single-root sequences. The presence of a list of root morphemes makes it possible to trace the repetition of such morphemes, functioning in the high-frequency zone, in the composition of low-frequency nouns. The given example represents the following data: the root 'mix' occurs with the same number of lexemes (5 units) in both the high-frequency zone and the low-frequency zone. The total frequency of all lexemes is 280 tokens – 272 units in the high-frequency zone and 8 in the low-frequency zone.

The third step is the transferring of all data on frequency of use obtained in the process of quantitative calculations, from absolute to relative, and then to percentages, since the text corpus "Automotive" exceeds the other two corpora by 100 thousand tokens. The article does not give the calculation procedure itself, but only the final results.

Of interest is not only the natural equalization of all data and reducing them to one type of value in order to obtain objective calculation results, but also a comparative analysis of absolute and relative values, which can help to better understand linguistic processes from a statistical point of view.

The fourth step is to present the results of the text research. Here it is advisable to indicate the following facts: 1) the number of root morphemes found in the texts of three specialties; 2) the number of lexemes formed on the basis of root morphemes; 3) the frequency with which lexemes with root morphemes are used, i.e. their repeatability.

In the high-frequency zone of three frequency dictionaries, 685 different roots were recorded, on the basis of which derivative nouns were formed. However, here it is advisable to take into account all productivity phenomena that occur in text corpora, so when making calculations it is better to also use the units recorded in the low frequency zone. The number of root morphemes in each list, their productivity and repeatability can be represented as follows.

Table 1

Root morphemes in the texts corpora and their (morphemes) percentage

Text corpora	Quantity		
	Root morphemes (%)	The number of different lexemes with root morphemes (%)	Frequency of use of lexemes with root morphemes (%)
Chemical engineering	353 (0,17–0,18%)	486 (0,24%)	30980 (15–16%)
Automotive	401 (0,13%)	599 (0,20%)	47189 (16%)
Electrical engineering	348 (0,17%)	471 (0,23–0,24%)	37452 (19%)

The calculations performed demonstrate that the share of root morphemes (in percentage terms) in the texts of the three specialties is almost the same in the text corpora "Chemical Engineering" and "Electrical Engineering" and somewhat less in the texts of the specialty "Automotive Engineering". Quantitative values for lexemes with root morphemes are repeated again in the specialties "Chemical Engineering" and "Electrical Engineering", and again somewhat less in the automotive industry texts (0.20%). That is, the calculated productivity of root morphemes in texts on chemical engineering and electrical engineering is the same, but in texts on automotive engineering it decreases by 0.004%.

And, finally, the frequency (repetition) of the use of lexemes with root morphemes is almost the same in quantitative terms in the "Chemical Engineering" and "Electrical Engineering" corpora. Then, for explanation, we can take absolute frequencies to clearly show the contradictory processes occurring in the linguistic phenomenon being described. So, if we take absolute frequencies, then in the "Automotive Engineering" corpus the frequency of use of lexemes with root morphemes exceeds, for example, by 17,000 units the number of lexemes with root morphemes, which is recorded in the "Chemical Engineering" corpus, and by 10,000 units the value that is presented for electrical engineering. And this is understandable, since the text corpus of the specialty "Automotive Engineering" itself has 100 thousand more tokens than the other two corpora.

However, the percentages (relative values) provide completely unexpected data: the frequencies of use of lexemes with root morphemes are almost the same in the specialties "Chemical Engineering" and "Automotive Engineering", although the values of the absolute frequencies, as already mentioned, are completely different. The relative magnitude of the functioning of root morphemes in the texts of the specialty "Electrical Engineering" is the highest.

Such a high occurrence of lexemes with root morphemes in the texts of the specialty "Electrical Engineering" indicates that the authors of new objects, inventions and developments described in the texts are trying to preserve the already available nomenclature of linguistic means, which are expressed, including in root morphemes, and not to introduce the new ones that are still practically unused, which may not be entirely clear to readers of their scientific articles. That is, the same root is the "core" for many derived lexemes.

And finally, the fifth step can be an example in which the results of the analysis of the most successful linguistic units in terms of their statistical characteristics are presented. In our case it is a description of the most productive root morphemes, as well as their frequency (repetition) values. The results of calculating the productivity and frequency of use of morphemes show that there are very few productive morphemes of this type in the text corpora under study. Thus, in the automotive corpus, only three productive morphemes are registered – 'act', 'form', 'press'. Each of them is the part of seven lexemes. For example, the root 'act' is used in the following lexemes: *'action', 'reactor', 'activity', 'actuator', 'reaction', 'double-acting', 'actuation'*. In electrical engineering texts, only one productive root morpheme was found – 'form', which creates eight lexemes, for example, *'form', transformer', 'autotransformer', 'performance', 'information', 'deformation', 'formation', 'transformation'*. The presence of such a large number of derivatives for one single morpheme fully confirms the conclusions made above regarding the saving of linguistic means by the authors when describing new developments in texts on electrical engineering. No productive root morphemes were found in chemical engineering texts.

We should note that all productive root morphemes are used in texts with high frequency which demonstrates the direct dependence of this statistical characteristic on the frequency of use.

Conclusions. From all of the above, the following conclusions can be drawn.

1. To analyze any linguistic phenomenon found in texts of any type of discourse, one should use not one text corpus, but several. The presence of several text corpora allows us to determine the general integral and differential characteristics of the object under study. Thus, in this article, the compilation of three text corpora based on texts in specialties that have different scientific topics made it possible to generalize the results obtained.

2. The research has shown that dividing the resulting probabilistic statistical models (frequency dictionaries) into high-frequency and low-frequency zones helps to obtain more accurate results, since the low frequency of occurrence of lexemes demonstrates the minimal probability of their use in texts.

3. To obtain results that are reliable from a statistical point of view, it is necessary to use text corpora of the same volume for analysis, or use relative values for further calculations, or present the final results of quantitative calculations in percentages. The latter is the most appropriate to show the distribution of values taking into account different sizes of text corpora.

4. When presenting analysis data in which the object of study is a very specific issue of linguistics, it is advisable to use all units of texts, i.e. operating both in the high-frequency zone and in the low-frequency zone.

5. As examples, it is best to use lexical objects in which statistical characteristics have the largest values. In this case, the idea of these objects will be the most complete.

Bibliography:

1. Koval, N., Grodska, E., Kokkina, L., Mardarenko, O., & Lebedeva, O. (2023). Comparative evaluation of grammatical phenomena among the different specialty students. *Amazonia Investiga. 12* (72). 91-100. DOI: https://doi.org./10.34069/AI/2023.72.12.8

2. Borisenko T. I., Tsinovaya M. V., Tsapenko L. E., Sirotenko T. V. The influence of component semantics in modal verb constructions with the modal verbs of "obligation" on their grammatical and statistical features (on the basis of the technical discourse texts). *Проблеми семантики слова, речення та тексту*. Київ: Київський національний університет ім.Тараса Шевченка. 2018. С. 82-100.
3. Ansarifar A., Shahriari H., & Pishghadam R. (2018). Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. *Journal of English for Academic Purposes*, 31, 58-71. http://dx.doi.org/10.1016/j.jeap.2017.12.008
4. Grodska Elina, Prokoichenko Anastasiia, Shapa Ludmila (2020). Cultural-linguistic approach to teaching and learning Spanish colour idioms and symbols. *Revista Romaneasca pentru Educatie Multidimensionala, 12 (1), Sup.1*. DOI: https://doi.org./10.18662/rrem/12.1 sup1/
5. Ludmila Shapa, Maria Nevreva, Marina Tsinovaya Statistics of low-frequency kernel (subordinating) models of the verbal word-groups in the text corpus "Radio electronics". *Одеський лінгвістичний вісник*. Одеса, 2018. С. 47-55.
6. Pochtaruk G. Ya., Zaitseva O. Yu., Moiseeva E. A., Sirotenko T. V. Comparative analysis of the semantic structure of the high-frequency word *unit* (on the material of scientific and technical discourse field "Automation of heat and power processes"). *Науковий ввісник МГУ.* Одеса, 2018. Вип. 33. С. 93-97.
7. Tsapenko L.E., Lebedeva E.V., Gvozd O.V. Implementation of verb forms in the texts of scientific and technical discourse. *Закарпатські філологічні студії*. Ужгород, Ужгородський національний університет, 2023. № 30. С. 128-133. http://zfs-journal.uzhnu.uz.ua/index.php/30-2023
8. Borisenko T. I., Kudinova T. I., Petrova E. I., Tomenko M. G. The role of the homogeneity/inhomogeneity criterion in determining the statistical reliability of a frequency dictionary (on the basis of the frequency dictionary "Radioelectronics"). *Закарпатські філологічні студії*. Ужгород: УДУ, 2019. Вип. 8. Т. 2. С. 164-168.
9. Shapa L. N., Tomenko M. G., Lebedeva E. V., Gvozd O. V. The influence of quantitative and qualitative features on statistical indicator of materiality/immateriality of discrepancies in the number of verbs in frequency dictionaries of different types. Закарпатські філологічні студії. *Закарпатські філологічні студії*. Ужгород: УДУ, 2018. Вип. 6. С. 162-169.
10. Неврева М. Н., Лебедева Е. В., Гвоздь О. В., Ершова Ю. А. Неврєва М. Н., Лебедєва Є. В., Гвоздь О. В., Єршова Ю. А. Англійська імовірнісно-статистична модель технічної спеціальності "Хімічне машинобудування" (частотний словник). *Науковий вісник МГУ*. Одеса, 2018. № 36. С. 71-75.
11. Неврева М. М., Дьяченко Г. Ф., Топча Н. І. Взаємозв'язок статистичних та лексичних особливостей іменних суфіксальних морфем у текстах наукового функціонального стилю. *Одеський лінгвістичний вісник*. Одеса, ОНЮА. 2017. С. 181-186.

Неврева М., Дуванська І., Мікешова Г. Послідовність процедур для отримання надійних результатів визначення статистичних характеристик лінгвістичних об'єктів

Анотація. Стаття представляє можливу послідовність процесу аналізу статистичних характеристик лінгвістичної одиниці, яка функціонує в текстовому корпусі. В якості об'єкта розглянуто статистичну характеристику продуктивності творення нових слів за допомогою дерівційних елементів, що додаються до кореневих морфем іменника. Матеріалом для аналізу є текстові корпуси трьох технічних спеціальностей, що відносяться до науково-технічного дискурсу – «Хімічне машинобудування», «Автомобільна техніка» та «Електротехніка». Вони складені на основі наукових статей, опублікованих у журналах відповідних галузей: Chemical Engineering Progress; Process Engineering; Machinery; Chemical and Process Engineering; Automobile Engineering, Auto Industry; IEEE Transactions on Power Apparatus and Systems, Proceedings of the Institution of Electrical Engineers. Розміри текстових корпусів «Хімічне машинобудування» та «Електротехніка» – 200 тис. слововживань, «Автомобільна техніка» – 300 тис. слововживань. Усі три області текстових корпусів мають різні спеціальності, не пов'язані між собою своєю тематикою. Це дає можливість отримати узагальнені результати, які дають змогу взяти за стилістичний маркер ту чи іншу інтегруючу чи диференційну ознаку. У статті детально описано послідовність необхідних кроків для отримання достовірних даних про продуктивність кореневих морфем іменника. У процесі дослідження застосовані такі методи: контекстний аналіз, за допомогою якого з текстів виділено всі похідні; статистичні методи розрахунку, які використовувалися при створенні ймовірнісно-статистичних моделей (частотних словників), переведення абсолютних величин у відносні; метод опитування досвідчених фахівців спеціальностей, взятих за матеріал для складання корпусів текстів тощо. Основним об'єктом дослідження є кореневі морфеми іменників, що зустрічаються в текстах з найбільшою частотністю для опису предмета, щоб результати аналізу біли якомога повнішими і очевиднішими.

Ключові слова: частота, продуктивність, значення (математичне), абсолютні та відносні частоти, текстовий корпус.