УДК 81'32
DOI https://doi.org/10.32782/tps2663-4880/2022.25.1.47

STATISTICAL METHODS FOR DETERMINING THE DIFFERENTIAL CHARACTERISTICS IN GRAMMAR PHENOMENA OF TEXTS (ON THE MATERIAL OF TEXT CORPORA OF TECHNICAL SPECIALTIES)

СТАТИСТИЧНІ МЕТОДИ ВИЗНАЧЕННЯ ДИФЕРЕНЦІЙНИХ ХАРАКТЕРИСТИК У ГРАМАТИЧНИХ ЯВИЩАХ ТЕКСТІВ (НА МАТЕРІАЛІ КОРПУСІВ ТЕКСТІВ ТЕХНІЧНИХ СПЕЦІАЛЬНОСТЕЙ)

Tsapenko L.E.,
orcid.org/0000-0002-5088-2393
PhD, Associate Professor,
Associate Professor at the Foreign Language Department
Odessa National Polytechnic University

Lebedeva E.V.,
orcid.org/0000-0002-5088-2393
Senior Lecturer at the Foreign Language Department
Odessa National Polytechnic University

Gvozd O.V.,
orcid.org/0000-0001-9882-8890
Senior Lecturer at the Foreign Language Department
Odessa National Polytechnic University

The article presents a description of statistical methods for determining units that function in the texts of different fields of knowledge, but related to the common scientific and technical discourse and having different quantitative values at the grammatical level. The object to which statistical calculations were applied was the aspect-temporal paradigmatic forms of finite verbs. Three text corpora, respectively, of three specialties – “Acoustics”, “Chemical Engineering” and “Automation of Heat and Power Processes” were used as the research material. Each corpus includes 100 thousand tokens each. The total sample size is thus 300 thousand tokens. The source for the material were articles from scientific journals published in England and the USA. Text corpora of specialties were used that are not thematically related to each other, which makes it possible to generalize the results obtained and determine some integral and differential characteristics inherent in the texts of scientific and technical discourse as a whole. The main research methods were the following: the method of structural and probabilistic analysis, contextual analysis and elements of distributive analysis, the method of expert assessment, the method of rank correlation. In addition a quantitative analysis of the studied units was used, explaining the causes of frequency and taking into account extralinguistic factors that determine some specific features of the scientific style. The analysis of the frequencies of occuring the aspectual-temporal paradigmatic forms of a finite verb in each text corpus and their comparison by the overall frequency makes it possible to trace the peculiarity of the paradigmatic forms-of-verb implementation in specific conditions. The authors have found that although the principle of the statistical parameter of frequency is rather rigidly preserved in the presented text corpora (meaning they belong to the same type of discourse in terms of the frequency of usage of the text units under consideration), nevertheless, we can talk about probable cases of discrepancy in the frequency of occurring of some of them in different text corpora included in the scientific and technical discourse. In our case, these are some forms of aspect-temporal verbal paradigm that have different statistical parameters in the texts of these technical specialties.

Key words: aspectual-temporal paradigmatic forms, frequency of occurring, scientific and technical discourse, token, contextual analysis.

Стаття представляє опис статистичних методів для визначення одиниць, що функціонують у текстах різних галузей знання, але належать до загального науково-технічного дискурсу та мають на граматичному різні кількісні показники. Об'єктом, до якого застосовувалися статистичні обчислення, було обрано видо-часові парадигматичні форми фінітних дієслів. Як матеріал дослідження були використані три текстові корпуси відповідно трьох спеціальностей – “Acoustics”, “Chemical Engineering” та “Automation of Heat and Power Processes”. Кожен корпус включав 100 тисяч слововжитків кожен. Загальний обсяг вибірки становить таким чином 300 тис. слововжитків. Джерелом для формування корпусів послужили статті з наукових журналів, виданих в Англії та США. Були використані текстові корпуси спеціальностей, які тематично не пов'язані між собою, що дає змогу узагальнити отримані результати та визначити деякі інтегральні та диференціальні характеристики, властиві текстам науково-технічного дискурсу загалом. Основними методами дослідження були такі: метод структурно-ймовірнісного аналізу, контекстуальний аналіз та елементи дистрибутивного аналізу, метод експертної оцінки, метод рангової кореляції. Крім того, використовувався кількісний аналіз досліджуваних одиниць з поясненням причин частотності та врахуванням екстралінгвістичних факторів, які детермінують деякі специфічні особливості наукового стилю. Аналіз частот вживання видо-часових парадигматичних форм фінітного дієслова в кожному текстовому корпусі та їх зіставлення за загальною частотою дає можливість простежити своєрідність реалізації парадигматичних форм

дієслова у конкретних умовах. Було встановлено, що хоча принцип статистичного параметра частотності досить жорстко зберігається у представлених текстових корпусах (мається на увазі їх віднесеність до одного типу дискурсу по частотності вживання текстових одиниць), проте можна говорити про можливі випадки розбіжності за частотою використання деяких з них в різних текстових корпусах, включених до науково-технічного дискурсу. У нашому випадку це деякі форми видо-часової дієслівної парадигми, які мають різні статистичні параметри в текстах зазначених технічних спеціальностей.

Ключові слова: видо-часові парадигматичні форми, частота вживання, науково-технічний дискурс, слововжитки, контекстуальний аналіз.

Formulation of problem. Analysis of the latest articles. One of the most frequently described types of discourse at present is the one that includes scientific and technical text corpora. Modern computer information processing tools make it possible to create corpora of almost any field of knowledge, which, of course, is very important for satisfying the needs of applied linguistics [5; 6; 7; 8; 9; 10].

But at the present stage of the development of linguistic science a simple representation of text sets is no longer enough. Researchers strive to get answers that can only be provided by the analysis of text units, their lexical, grammatical and other characteristics that give a clear understanding of their functioning. The results of such an analysis allow not only to draw conclusions about the current state of scientific discourse, but also to anticipate possible changes in future scientific works.

In addition some problems remain in discourse studies that have not been resolved in previous studies i.e. those that until recently remained outside the field of attention of researchers. This concerns, first of all, such a question as the differentiation of text corpora of various technical fields at the grammatical level. The reason for this situation can be considered the conviction of scientists that the belonging of a text to one or another technical (or non-technical) specialty can only affect lexical-semantic features, but not its grammatical characteristics [1; 2; 3; 4].

However now linguists using statistical data obtained as a result of a survey of text corpora [11; 12; 13; 14] note certain quantitative differences in the implementation of grammatical units in the texts belonging to scientific and technical discourse but referring to various specialties. For example, there is a different frequency of use of some syntactic and morphological units. Thus, along with specific linguistic phenomena that are characteristic for all texts of the same type of discourse and which are used with almost the same frequency in these texts, there is a limited number of units (language means), the statistical parameters of which may differ significantly in text corpora of various specialties, although belonging to the same common type of discourse.

Formulation of task. The purpose of the article is to describe statistical methods for determining units that function in the texts of various fields of knowledge, but referred to scientific and technical discourse and having different quantitative values at the grammatical level. The aspectual-temporal paradigmatic forms of finite verbs were chosen as the object to which statistical calculations were applied.

Basic material presentation. Three text corpora, respectively, of three specialties – "Acoustics", "Chemical Engineering" and "Automation of Heat and Power Processes" are used as research material. Each corpus includes 100 thousand tokens each. The total volume of the material is thus 300 thousand tokens. Articles from scientific journals published in UK and the USA served as a source for the compilation of corpora. The author should stress that the use of text corpora dealing with specialties that are not thematically connected to each other makes it possible to generalize the results obtained and determine some integral and differential characteristics inherent in the texts of scientific and technical discourse as a whole.

The following methods were used: the method of structural and probabilistic analysis, contextual analysis and elements of distributive analysis, the method of expert assessment, the method of rank correlation. In addition, a quantitative analysis of the studied units was used, explaining the causes of frequency and taking into account extralinguistic factors that determine some specific features of scientific and technical discourse.

The topicality of the presented study is due, on the one hand, to the functional significance of the considered grammatical phenomenon in language and speech, on the other hand, to the insufficient knowledge of the features of its functioning in the engineering texts. The authors try to prove that the assignment of texts of different profiles or specialties to the same type of discourse does not mean that the frequency of the use of certain morphological and syntactic phenomena in these texts should be the same (or approximately the same). It can vary within the same discourse depending on the types or genres of texts which can differ significantly from each other in terms of the degree of abstraction, information content, and specific goal setting.

Thus, the present work is in some contradiction with the already generally accepted statements that only statistical parameters can serve as a sufficient

basis for including the exact text corpora in this or that type of discourse [1; 2; 3; 4].

As already mentioned in order to be able to judge the integral and differential features of different texts of this discourse, and in particular, texts referred to different areas of scientific and technical knowledge, the text corpora of three specialties have been created: "Acoustics", "Chemical Engineering " and "Automation of Heat and Power Processes".

The homogeneity of the text corpus is ensured by the belonging of all the texts included in it to the common discourse (scientific) and one genre (scientific article).

The content of individual articles of each corpus was determined in accordance with the specific share and number of subject areas of the semantic space of each specialty, determined by interviewing specialists in these fields of technology when compiling frequency dictionaries of these specialties [5; 7].

The sufficiency of the sample size was checked according to the formula adopted in linguistics:

$$N = \frac{Sp^2}{\varepsilon^2 \cdot f}, \qquad [1]$$

where N – is the minimum size for obtaining the reliable information of the total text corpus (in word forms);

f – is the relative frequency of the analyzed grammatical phenomena (the number of occurrences divided by the number of words in the text corpus);

ε – relative error;

Sp – is a coefficient used in linguistic research, equal to 1.96 [1].

When examining the text sets (corpora) "Acoustics", "Chemical Engineering" and "Automation of Heat and Power Processes", it was found that the relative frequency (f) of the aspect-temporal paradigmatic forms of a finite verb is 0.06 since the size of one continuous text totality is 100 thousand tokens. Therefore, out of 25 journal articles on which the content of each corpus of three specialties is based, 6126 5607; 6084 finite forms of verbs were selected respectively. The relative error with such a size of each corpus is equal to:

$$\varepsilon = \frac{1,96}{\sqrt{N \cdot f}} = \frac{1,96}{\sqrt{100000 \cdot 0,06}} = 0,025 = 2,5\%$$

The obtained data (f, ε) are substituted into the formula

$$N = \frac{z\sigma^2}{\varepsilon^2 \cdot f} = \frac{1,96^2}{0,025^2 \cdot 0,06} = 96000$$

In other words the selected number of examples at ε = 2.5% covers 97.5% of the text, i.e. with an accuracy of 97.5% it ensures the reliability of statistical characteristics in a text corpus with a size of 96 thousand words.

At present it is commonly accepted that the total corpus is considered quite sufficient if it covers 70–60% of the entire set of texts and the relative error in the analysis of linguistic phenomena is taken in the range from 3% to 25% [1; 6].

The units of counting in our research were the paradigmatic forms of the verb-predicate.

The selected elementary sentences were analyzed from the point of view of the implementation of voice and aspectual-temporal paradigmatic forms of the finite verbs of the indicative mood and their structure. Thus each verb was endowed with a certain set of characteristics to be calculated.

After processing each text corpus the tables were compiled, then summary tables and generalized lists of fixed units in descending order of frequencies were compared both within the same specialty and between different specialties, i.e. the method of correlation of signs and rank correlation were used.

The identified aspectual-temporal paradigmatic forms were divided according to their frequencies into high-frequency ones (which occurred more than 100 times in the texts of these specialties) and low-frequency ones. All the aspectual-temporal paradigmatic forms of the finite verb with low frequencies were combined into one group of "low-frequency" forms for the convenience of calculations and to ensure statistical reliability.

Using the quantitative and qualitative parameters of the verbal paradigm implemented in the texts of the studied scientific fields of knowledge, and identifying high-frequency and low-frequency aspectual-temporal forms, we studied the syntactic conditions for the functioning of only high-frequency aspect-temporal paradigmatic forms of the finite verbs of the active and passive voices, which show the specificity of "behavior" of the morphological categories of the verbs in finite forms in the texts of different fields of scientific discourse.

The peculiarity of scientific text is generally determined by the purpose for the most economical and at the same time the most accurate presentation. This factor, which lies outside the actual linguistic sphere, to a large extent affects the syntactic characteristics, the use of certain grammatical classes of words, and the frequency of their occurrence in the text.

Since the verb can be combined with nouns, adjectives, adverbs, infinitives, gerunds the behavior

of high-frequency aspectual-temporal paradigmatic forms of finite verbs was studied in terms of their immediate environment on the right (words and phrases), associated with the studied verbs by a syntactic or semantic links.

Almost 50 syntactic constructions were registered with verbs in the active voice, with verbs in the passive voice – only 20 constructions; for verbs in the active and passive voice, lexical units were identified that are characteristic of these syntactic constructions. Only high-frequency syntactic constructions were used for the further research.

The statistical approach to the study of text involves the use of a certain mathematical apparatus, the purpose of which is to obtain objective data on the discrepancy between the frequencies of language units in the same array or in different arrays of texts.

In this work the following mathematical calculations have been carried out. When performing the statistical analysis for each aspect-temporal paradigmatic form, the average frequency was calculated for the 1000 token text set according to the formula

$$\bar{x} = \frac{a}{n}\sum_{1}^{n}\frac{x_i}{a_i} \qquad [15]$$

where a – is the total text set size

n – the number of text sets

x_i – is the absolute frequency of the given shape

a_i – is a size of each text (set).

Then after calculating the average frequency the standard deviation (σ) and the value of the change in the average frequency ($\sigma_{\bar{x}}$) were found using the formulas:

a) $$\sigma = \sqrt{\sum\frac{(x_i = \bar{x})^2 n_i}{N}}$$

where the $(x_i = \bar{x})$ – is the difference between each of the absolute and average frequencies;

n_i – is the part of the text set with absolute frequency

N – the amount of the researched text corpora

b) $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

In addition, using the value of fluctuations in the average frequency ($\sigma_{\bar{x}}$), a relative research error was revealed for each aspectual-temporal paradigmatic form:

$$\varepsilon = \frac{1{,}96 \cdot \sigma\bar{x}}{\bar{x}} \cdot$$

The comparison of the percentages of the frequencies of the compared values was carried out according to the formula

$$t = \frac{P_1 - P_2}{SEd\%} \qquad [16, p. 61]$$

where P_1 – is the percentage indicator of the first group;

P_2 – is the percentage indicator of the second group.

$$SEd\% = \sqrt{PQ(\frac{1}{n_1} + \frac{1}{n_2})}, \quad P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

where P – is arithmetical mean of the percentage indicator of the both groups:

$$Q = 1 - P$$

n_1 – is the absolute feature frequency in the first group;

n_2 – is the absolute feature frequency in the compared group абсолютная;

The number of degrees of freedom is determined by the formula:

$$f = (n_1 - 1) + (n_2 - 1),$$

where n_1 – is the number of text corpora compiled for one specialty;

n_2 – is the number of text corpora compiled for the compared specialty.

The degree of freedom indicator is used to determine the critical value of the Student's criterion according to the table. Values above the critical point indicate significant frequency discrepancies.

The correlation analysis was carried out according to the formula:

$$c = \frac{6\sum d^2}{N(N^2 - 1)} \qquad [17, p. 8]$$

where c – is the rank correlation coefficient

d – is the difference between the frequency ranks of two features in one set of texts

N – the number of rows (features).

To assess the nature of frequency fluctuations we calculated (for each paradigmatic form of the verb) the coefficient of variation (V)

$$V = \frac{V}{\bar{X}} \cdot 100 \qquad [17, p. 36]$$

which shows the ratio of the standard deviation to the average frequency expressed as a percentage. The critical threshold of the coefficient of variation is 40%, above which the frequency variation in the text corpora is not random but regular.

Conclusions. The analysis of the frequencies of the use of aspectual-temporal paradigmatic forms of finite verbs in each text corpus and their comparison by the common frequency makes it possible to trace the originality of the implementation of the paradigmatic forms of the verb in specific conditions. It was found that although the principle of the statistical parameter of frequency is rather rigidly preserved in the presented text corpora (meaning they belong to one type of discourse in terms of the frequency of use of the text units under consideration), nevertheless, the value of the coefficient of variation – 40% – is high enough to be able to talk about probable cases of discrepancy in the frequency of use of some of them in different text corpora included in the scientific and technical discourse. In our case these are some forms of the aspectual-temporal verbal paradigm that have different statistical parameters in the texts of the above technical specialties.

The study of modern works devoted to this issue has led to the conclusion that the best for research of this nature is a combination of methods of continuous text analysis, structural and probabilistic analysis, elements of distributive analysis, comparative analysis at the morphological and syntactic levels using mathematical methods to establish the reliability of the results obtained. Such kind of combination allows the most complete and comprehensive presentation of the results of text corpora analysis, which is carried out in order to detect the integral and differential characteristics of a particular linguistic phenomenon that has manifested itself in texts of various specialties referred to the same type of discourse.

BIBLIOGRAPHY:

1. Piotrovsky Rajmund G. Quantitative Linguistics. *An International Handbook. Walter de Gruyter*. Berlin. New-York. 2005. 1027 p. [edited by Reinhard Köhler, Gabriel Altmann]

2. Alekseev P. M. Statistical lexicography (typology, compiling and application of frequency dictionaries). 1975. 120 p.

3. Andreev N. D. Statistical-combinatorial methods in theoretical and applied linguistics. 1967. 404 p.

4. Zakharov V.P. Corpus linguistics: teaching method [student teaching book]. 2005. 48 p.

5. Dyachenko G. F. Guidelines for working with English special vocabulary for students of the specialty “Acoustics” (minimum frequency dictionary). Odessa, 1985. 60 p.

6. Перебійніс В. І., Муравицька М. П., Дарчук Н. П. Частотні словникі та їх використання. Київ, 1985. 202 р.

7. Tomasevich N.P. Terminological vocabulary of the English sublanguage of the automotive industry and its interaction with other lexical layers: author. diss. … cand. philological sciences: spec. 10.02.04 “Germanic languages”. Odessa, 1984. 16 p.

8. Shapa LN Forms and functions of adjectives in the scientific and technical text (on the material of the English sublanguage of Power Supply): diss. … cand. philol. Sciences: 10.02.04. Odessa, 1990. 201 p.

9. Melnikova M.V. English-Pussian Dictionary of word combinations and cliché for a specialist-researcher. Publ. PSPU, Perm’. 2000. 272 p.

10. Benson M., Benson E., Ilson R. The BBI combinatory dictionary of English: a guide to word combinations. Amsterdam – Philadelphia, 1997 [a companion volume to the Lexicographic Description of English].

11.Циновая М. В. Формы и содержание синтаксических конструкций с глаголом may/might в текстах научной коммуникации. *Науковий вісник Міжнародного гуманітарного університету. Серія «Філологія».* Одеса,2015. № 14. Ч.2. С. 92–96.

12. Tsinovaya M. V. Lexical component of the second constituent of modal verb constructions in the texts of scientific-technical communication. *Вісник харківського національного університету імені В. Н. Каразіна Серія “Романо-германська філологія. Методика викладання іноземних мов”*. Харків, 2014. № 1102. С. 155–159.

13. Борисенко Т. И., Кошуба М.В., Мардаренко Е.В., Циновая М.В. Особенности функционирования модальных глагольных конструкций в подъязыках техники. *Записки з романо-германської філології*. Одеса, 2014. Вип. 1(32). С. 25–34.

14. Рогачева Н. В. Морфемна структура іменників у текстах різних функціональних стилів (на матеріалі англійської мови): Автореф. дис... канд. філол. наук: 10.02.04 “Германские языки”. / Київський держ. лінгвістичний ун-т. К., 2000. 25 с.

15. Nosenko I.A. The beginning of statistics for linguists. 1981. 65 p.

16. Бровченко Т.О. Словесний наголос в сучасній українській мові. Київ, 1969. 168 стор.

17. Golovin B.N. Experience of using correlation analysis in language learning. *Вопросы статистической стилистики*. Киев, 1974. P. 5–16.