УДК 81’373.611
DOI https://doi.org/10.32782/tps2663-4880/2025.40.2.29

STATISTICAL FEATURES OF ENGINEERING TEXTS

СТАТИСТИЧНІ ОСОБЛИВОСТІ ТЕХНІЧНИХ ТЕКСТІВ

Nevreva M.M.,
orcid.org/0000-0001-9923-1041
Candidate of philological Sciences, PhD,
Associate Professor at the Foreign Languages Department
National University “Odesa Polytechnic”

Prysyazhnyuk O.A.,
orcid.org/0000-0001-7357-5516
Candidate of philological Sciences, PhD,
Associate Professor at the Foreign Languages Department
National University “Odesa Polytechnic”

Vorobyova K.V.,
orcid.org/0000-0003-3474-502x
Senior Lecturer at the Foreign Languages Department
National University “Odesa Polytechnic”

The presented article describes statistical data that characterize probabilistic-statistical models of three technical specialties. The material was text corpora of the specialties chemical engineering, automotive engineering and electrical engineering. The text corpora were compiled on the basis of scientific articles published in the corresponding journals in the USA and Great Britain. At the beginning, the authors assumed that since all three text corpora relate to scientific and technical discourse and do not interact in terms of general scientific topics, but represent completely different branches of technology, they could create an example of style-distinguishing marker in the course of the procedure of comparison. However, different approaches to the formation of text corpora, and, above all, the principles and goals of creating the semantic space of all three specialties led to a sharp difference between the quantitative values of all three frequency dictionaries. The authors of the probabilistic-statistical model of chemical engineering sought to evenly distribute text units across the frequency dictionary and, for this purpose, selected texts that had practically common technical problems, and, accordingly, repeated lexical units that described these problems. This was easily achieved by creating a simple and fairly limited scheme of semantic space, including a few problems that relate to the subjects of this specialty. The same was done by the authors of the text corpus of automotive engineering. As to the authors of the text corpus of electrical engineering they attempted to include practically all scientific and technical problems of this specialty in the semantic space. Therefore, firstly, within the frequency dictionary of electrical engineering itself, sharp differences in quantitative values were observed between the high-frequency and low-frequency zones. Secondly, when comparing the quantitative values of three probabilistic-statistical models, which were expected to be practically identical, the data of the dictionary of electrical engineering were completely different from the dictionaries of chemical engineering and automobile engineering. As a result of the study, the authors proposed to observe one of two conditions when forming text corpora: if it is supposed to consider as many technical problems as possible, then it is necessary to either significantly increase the size of text corpus, or to approach the selection of texts more carefully.

Key words: semantic space, probabilistic-statistical model, token, frequency dictionary, frequency zone.

У представленій статті описано статистичні дані, що характеризують ймовірнісно-статистичні моделі трьох технічних спеціальностей. Матеріалом слугували текстові корпуси спеціальностей хімічного машинобудування, автомобілебудування та електротехніки. Текстові корпуси були складені на основі наукових статей, опублікованих у відповідних журналах США та Великої Британії. Спочатку автори припускали, що оскільки всі три текстові корпуси стосуються науково-технічного дискурсу та не взаємодіють з точки зору наукових тем, а представляють абсолютно різні галузі техніки, вони можуть створити приклад стилістичного розрізнювального маркера в ході процедури порівняння. Однак, різні підходи до формування текстових корпусів, і, перш за все, принципи та цілі створення семантичного простору всіх трьох спеціальностей, призвели до різкої різниці між кількісними значеннями всіх трьох частотних словників. Автори ймовірнісно-статистичної моделі хімічного машинобудування прагнули рівномірно розподілити текстові одиниці по частотному словнику та для цього відбирали тексти, що мали практично спільні технічні проблеми, і, відповідно, повторювані лексичні одиниці, що описували ці проблеми. Цього було легко досягти, створивши просту та досить обмежену схему семантичного простору, що включає кілька проблем, які стосуються проблем цієї спеціальності. Те ж саме зробили й автори текстового корпусу автомобілебудування. Що ж до авторів текстового корпусу електротехніки, то вони намагалися включити практично всі науково-технічні проблеми цієї спеціальності до семантичного простору. Тому, по-перше, в межах самого частотного словника електротехніки спостерігалися різкі відмінності в кількісних значеннях між високочастотною та низькочастотною зонами. По-друге, при порівнянні кількісних значень трьох ймовірнісно-статистичних моделей, які очікувалися практично ідентичними, дані словника з електротехніки повністю відрізнялися від значень словників хімічного та автомобільного будування. В результаті дослідження автори запропонували дотримуватися однієї з двох умов при формуванні текстових кор-

пусів: якщо передбачається врахування якомога більшої кількості технічних проблем, то необхідно або значно збільшити розмір текстового корпусу, або ретельніше підійти до відбору текстів.

Ключові слова: семантичний простір, ймовірнісно-статистична модель, слововживання, частотний словник, частотна зона.

Problem statement. **Literature review**. The presentation of statistical data in linguistics plays a decisive role in determining the type of discourse, i.e. classifying a text as belonging to a particular functional style. It is quantitative values that are considered the main reasons for including texts in a text collection [1; 2]. Why is it so important to classify a text as referring to a particular type of discourse? First of all, because computer translation is much easier if the text is a part of, for example, a scientific and technical discourse [3]. Such types of texts are usually so formalized, structurally identical [4], devoid of emotional coloring and, consequently, of a large number of idiomatic phrases, i.e. they use a strictly limited amount of data from certain levels of language, are subjected to the semantic inventory of sublanguages, etc., that a computer can translate such a text much more easily, i.e. with a minimum number of errors [5; 6; 7].

However, it is not only a matter of the possibilities of computer translation, but also of achieving the didactic goals, since corpus linguistics is at the intersection of the tasks of theoretical and applied linguistics. Since the applied linguists (teachers, translators, etc.) use computer corpora in teaching languages and to solve their professional problems it has been showed on the basis of specific systems implemented on a machine how the principle of targeted understanding of a language text is embodied in practice [8; 9].

Task Statement. The purpose of the article is to consider and describe the statistical data characterizing probabilistic-statistical models of specialties, the texts of which are referred to scientific and technical discourse, and also to explain the reasons for obtaining such data.

Base material. In order to obtain the results that could substantiate certain differential and integral characteristics of the compiled probabilistic-statistical models [10], three frequency dictionaries were considered – on chemical engineering, automotive engineering and electrical engineering. It is noteworthy that all three specialties do not have common theoretical themes, which may possibly make the results of the analysis the style-distinguishing markers for any text corpora referred to scientific and technical discourse.

Along with the description of quantitative parameters of the frequency dictionaries presented in this article, it was decided to take as an example some (best of all – the most frequent one, in order to operate with the reliable values) component of the dictionaries. For the description, the nouns functioning in the text corpora of the mentioned specialties were chosen.

So, let's describe first one of the dictionaries. The entire frequency dictionary, i.e. the probabilistic-statistical model of the specialty chemical engineering, consists of 6589 words. In order to carry out the research with statistically reliable data, the entire list was divided into two zones – high-frequency and low-frequency. The high-frequency zone includes the first 2000 words, which, being the most frequently used, cover almost 89% of the examined texts. These words represent the core of the probabilistic-statistical model of texts in this specialty, normally called the "base dictionary". The remaining 4589 words are assigned to the low-frequency zone. Both other dictionaries (automotive and electrical engineering) were similarly also divided into a base dictionary, including the 2000 most frequent words, and the rest, in which words with low frequency of occurrence function.

The units of the base dictionary of automotive engineering (two thousand words) cover 83% of the text corpus. The low-frequency zone includes 5037 text units.

The base dictionary of the electrical engineering specialty, i.e. the first 2000 words, covers 79% of the text corpus. The low-frequency zone includes 6308 words.

If we analyze the presented statistical data, we can state the following. The lowest coverage of the text corpus is for the electrical engineering specialty, the highest is for chemical engineering, and the automotive engineering specialty occupies an intermediate position. In addition, the list of the base dictionary ends at number 9 in the chemical engineering specialty, and at the automotive and electrical engineering specialties at number 7.

How can such a severance in the data values on coverage extent in the three described text corpora be explained? Here it is necessary to mention such a parameter necessary for corpus linguistics as the formation of the semantic space of the specialty, which covers technical topics describing the main phenomena and objects of the specialty. In the chemical engineering model, a more uniform distribution of vocabulary units is presented, the reason for which is undoubtedly the high concentration of

texts of similar topics included in the text corpus. We can conclude that the author sought to find texts in journals of the corresponding specialty that would record practically the same lexes, i.e. practically the same research problems. The text corpora of automotive engineering and especially electrical engineering practically lack such uniformity, since the authors of these probabilistic-statistical models tried to cover as many thematic problems of their specialties as possible.

Let us consider the low-frequency zone of these three frequency dictionaries, which, according to many researchers, is not only a potential reserve for future most frequent text units, but also a place where most terms are concentrated. The number of words included in this zone in the dictionary of chemical engineering is 4589; in the dictionary automotive engineering – 5037 units; in the dictionary electrical engineering – 6308 units. Here again we can talk about the degree of uniformity of text unit distribution, but not only. It is obvious that most terms in the texts chemical engineering have moved to the base list (this is also evidenced by the number 9 at the end of the basic dictionary). The same can be said about the terminological units of the automotive engineering specialty, since the number of words in the low-frequency list differs only slightly from the list in the dictionary of chemical engineering (only by almost 400 words). But the dictionary of electrical engineering has a significant difference in this matter – almost one and a half times the difference from the two mentioned dictionaries. Such a sharp severance in quantitative values between the high-frequency and low-frequency zones is a feature of a certain dissonance in the selection of theoretical and applied problems that the authors found in the studied journal articles. That is, if the authors of the dictionaries of chemical engineering and automotive engineering have adhered to a certain strategy of accumulating text resources and sought to record works dealing only with theoretical or only applied objects, then the author of the dictionary of electrical engineering tried to cover as many works as possible in their nature. Therefore, most text units have moved to the low-frequency zone. Theoretical scientists specializing in the fields of corpus, computer and structural linguistics [11; 12; 13] recommend adhering to an average strategy when selecting texts for a text corpus, which is demonstrated in the other two frequency dictionaries. This does not mean that the dictionary on electrical engineering has no scientific or applied value, on the contrary, it is a kind of model and one of the illustrative examples for young scientists who are going to form probabilistic-statistical models of various types of speech for all purposes.

Now let us consider the place of nouns in the probabilistic-statistical models of the three specialties. In each of the three basic dictionaries, nouns dominate over other parts of speech. They account for 44–56% of all registered words. Many scientists have paid attention to this characteristic of scientific and technical text, which is typical for different specialties. The reason for the nominal nature of scientific and technical discourse is the high information content of nouns [14]. The nominative function of a noun makes it especially important in describing various technical objects and phenomena. Since a fairly large number of nouns belong to the terminological layer of vocabulary, linguists argue that the main terminological fund consists of nouns.

The number of nouns in the basic dictionary of each specialty is as follows: in the dictionary of chemical engineering – 955 words (i.e. almost 50% of the entire base dictionary) and their absolute frequency fluctuates from 811 to 9; in the dictionary of automotive engineering – 1025 nouns (more than 50%), their absolute frequency fluctuates from 2126 to 7; in the dictionary of electrical engineering – 883 words (46%) with an absolute frequency from 3164 to 7.

The presented quantitative values allow us to draw the following conclusions. Firstly, they fully confirm the opinion of linguists about the dominance of nouns over other parts of speech in the texts of scientific and technical discourse (in our case, in the base dictionaries of specialties). Secondly, the statistical data once again testify to the fundamental difference in the selection of texts by the authors of these three probabilistic-statistical models. If we pay attention to the huge difference between the maximum values of nouns in the base dictionaries (in chemical engineering the maximum frequency is 811, in automobile engineering 2026, i.e. more than twice as much, in electrical engineering 3164) with almost the same number of words (in electrical engineering even less than in the other two dictionaries), then, of course, the content of the text corpus of the specialty chemical engineering from the point of view of the uniformity of the text distribution on practical problems of this specialty can be considered the most successful. The almost three times greater maximum value of nouns in the dictionary on electrical engineering shows that the terms with a high frequency of use are presented in the base dictionary (for example, current F=3164, voltage F=1465, frequency F=482 load F=360, phase F=334, etc.). They are few because most of them, as already indicated, are concentrated in the low-frequency zone. As in the previously presented other quantitative values, the maximum value of the

noun list in the dictionary on automotive engineering occupies an intermediate position.

Next, we calculate the percentage of coverage of these lists of nouns of these specialty base dictionaries: 955 nouns in the base dictionary of chemical engineering have a total absolute frequency of 51,641 tokens, the percentage of coverage of the entire text corpus is 26%.

As already mentioned, the base dictionary of automotive engineering contains 1,125 nouns. The cumulative frequency of these nouns is 67,889 tokens, while the coverage of the corpus texts by these nouns is 29%.

The frequency dictionary of electrical engineering in the high-frequency zone, i.e. the basic dictionary, contains 883 nouns. In the examined texts, they are represented by 53,125 tokens, which covers the texts by 27%.

It can be seen that the percentage of coverage of text corpora by nouns in all three specialties is almost the same. The automotive engineering dictionary is slightly higher, which is natural because it has a larger number of words. So, this phenomenon cannot be explained by any special reasons. But what attracts attention is the quantities of nouns in the electrical engineering dictionary. Having the smallest number of words (883 words) compared to the lists of the other two specialties, the nouns in this text corpus are in no way inferior to, for example, the level of coverage of texts by nouns in chemical engineering, but even exceed it (53125 and 51641, respectively). This is explained by the fact that the majority of the most frequent words-terms expressed by nouns (examples are given above) are in the high-frequency zone of the list.

Let's consider another quantitative values – the percentage of coverage of the low-frequency zone of the list by nouns. The value in the three described basic dictionaries turned out to be somewhat unexpected: 3% for the specialty chemical engineering and 2% each for automotive engineering and electrical engineering. Of course, for the specialty chemical engineering this value should be higher than in the other dictionaries, since the low-frequency zone here began with the number 8, but still it turned out to be almost the same everywhere. Although the number of words in this zone for the specialty electrical engineering is significantly higher than in the other two dictionaries, this does not affect the number of nouns on their level of coverage, which turned out to be the same as in the dictionary chemical engineering and automotive engineering.

Conclusions.

1. Despite the fact that the probabilistic-statistical models under consideration belong to the same type of discourse – scientific and technical – the quantitative data of one of the three dictionaries differ significantly from the other two (chemical engineering and automotive engineering), which can serve as a basis for further study of such sharp differences.

2. It was determined that significant differences in the numerical values of one of the three frequency dictionaries from the ones (values) of the other two frequency lists under consideration are caused by the lack of uniformity of the distribution of values in the dictionary of the electrical engineering specialty. Such a phenomenon is usually observed if the author, when compiling the text corpus, tries to cover as wide a range of topics and problems of the specialty as possible.

3. The recorded differences demonstrate the following conclusions of the authors: if the authors of frequency lists prefer uniformity of distribution of values to a greater coverage of technical problems of a particular specialty, then they can expect sharp jumps in values, opposition of high-frequency and low-frequency indicators, impossibility of performance of a comparative analysis with other statistical objects and other researching procedures. Therefore, it seems that the only alternative between the desire for uniformity of distribution of values (and at the same time loss in thematic problems of specialties) and the desire to cover as many problems of a particular specialty as possible is the fulfillment of at least one of two parameters: a) a significant increase in the volume of the text corpus; b) a more careful selection of texts and special attention to the terms used in the texts that describe a particular problem or object of study.

REFERENCES:

1. Dascal M. Pragmatics and the Philosophy of Mind. Vol. I: *Thought in Language*. John Benjamins Publishing Company. 1983. 207 p.

2. Sperber and WIlson (1986) Relevance: Communication and Cognition. May 1989. *Mind & Language* 4(1–2): 138–146. DOI: 10.1111/j.1468-0017. 1989.tb00246.x

3. Lu Y., Mei Q., Zhai C. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*. 2011. Vol. 14, no. 2. Pp. 178–203.

4. Riedl M., Biemann C. TopicTiling: A text segmentation algorithm based on LDA. *Proceedings of ACL 2012 Student Research Workshop*. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. Pp. 37–42.

5. Chuang J., Gupta S., Manning C., Heer J. Topic model diagnostics: Assessing domainrelevance via topical alignment. *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. Ed. by S. Dasgupta, D. Mcallester. Vol. 28. JMLR Workshop and Conference Proceedings, 2013. Pp. 612–620.

6. Jacksi K., Dimililer N., Zeebaree S. R. M. A survey of exploratory search systems based on LOD resources. *Proceedings of the 5th International Conference on Computing and Informatics,* ICOCI 2015. School of Computing, Universiti Utara Malaysia, 2015. Pp. 501–509.

7. Mimno D., Li W., McCallum A. Mixtures of hierarchical topics with pachinko allocation. ICML. 2007.

8. Hoffman M. D., Blei D. M., Bach F. R. Online learning for latent Dirichletallocation. *NIPS*. Curran Associates, Inc., 2010. Pp. 856–864.

9. Koltsov S., Koltsova O., Nikolenko S. Latent Dirichlet allocation: Stability and applications to studies of user-generated content. *Proceedings of the 2014 ACM Conference on Web Science.* WebSci'14. New York, NY, USA: ACM, 2014. Pp. 161–165.

10. Vulic I., De Smet W., Tang J., Moens M.-F. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. *Information Processing & Management*. 2015. Vol. 51, no. 1. Pp. 111–147.

11.Zuo Y., Zhao J., Xu K. Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*. 2016. Vol. 48, no. 2. Pp. 379–398.

12. Yin Z., Cao L., Han J., Zhai C., Huang T. Geographical topic discovery and comparison. *Proceedings of the 20th international conference on World wide web.* ACM. 2011. Pp. 247–256.

13. Andrzejewski D., Zhu X. Latent Dirichlet allocation with topic-in-set knowledge. *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*. SemiSupLearn '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. Pp. 43–48.

14. Potter S. Changing English. Andre Deutsch, Second reversed edition. 1975. 306 p.