РОЗДІЛ 9
СТРУКТУРНА, ПРИКЛАДНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

УДК 81’32

THE ROLE OF THE HOMOGENEITY/INHOMOGENEITY CRITERION IN DETERMINING THE STATISTICAL RELIABILITY OF A FREQUENCY DICTIONARY (ON THE BASIS OF THE FREQUENCY DICTIONARY “RADIOELECTRONICS”)

РОЛЬ КРИТЕРІЯ ОДНОРІДНОСТІ/НЕОДНОРІДНОСТІ ПІД ЧАС ВИЗНАЧЕННЯ СТАТИСТИЧНОЇ НАДІЙНОСТІ ЧАСТОТНОГО СЛОВНИКА (НА ПРИКЛАДІ ЧАСТОТНОГО СЛОВНИКА «РАДІОЕЛЕКТРОНІКА»)

Borisenko T.I.,
orcid.org/0000-0002-9839-1761
PhD, Associate Professor,
Associate Professor of the Foreign Language Department
of the Odessa National Polytechnic University

Kudinova T.I.,
orcid.org/0000-0003-3302-6699
Senior Lecturer of the Foreign Language Department
of the Odessa National Polytechnic University

Petrova E.I.,
orcid.org/0000-0003-0637-0806
Senior Lecturer of the Foreign Language Department
of the Odessa National Polytechnic University

Tomenko M.G.,
orcid.org/0000-0002-4149-0320
Senior Lecturer of the Foreign Language Department
of the Odessa National Polytechnic University

The article discusses one of the statistical linguistics criterion – homogeneity/inhomogeneity one, the use of which promotes compiling the representative text corpora followed by the forming of statistically reliable frequency dictionaries (in the presented article it is the one of Radio electronics). This criterion was applied in comparing the seven frequency dictionaries referred to scientific and engineering discourse, as well as to discourses of other types. The following characteristics of the homogeneity/inhomogeneity criterion were considered: “Chronological homogeneity”, “Discursive homogeneity”, “Homogeneity of a sublanguage”, “Thematic homogeneity of texts”, because, according to the authors’ opinion, they (characteristics) are of the basic practical importance in creating frequency dictionaries. The text corpora needed for the study were formed on the basis of scientific articles of the relevant specialties journals of the USA and Great Britain, scientific and technical documentation, newspaper articles. The results of comparative analysis of seven text corpora (including Radio electronics) possessing different quantitative and qualitative peculiarities by the homogeneity/inhomogeneity criterion (its four characteristics), have showed the statistical data that completely reflected the individual features of each of the text corpora.

Key words: discourse, period of time, representativeness, sublanguage, text corpus.

У статті розглядається один із критеріїв лінгвостатистики – критерій однорідності/неоднорідності, використання якого сприяє створенню репрезентативних текстових корпусів із наступною компіляцією статистично надійних частотних словників (у статті це – частотний словник із радіоелектроніки). Даний критерій перевірявся шляхом порівняння семи частотних словників, які відносяться до науково-технічного дискурсу, а також до дискурсів інших типів. Розглядалися такі характеристики критерію однорідності/неоднорідності: «хронологічна однорідність», «дискурсивна однорідність», «однорідність підмови», «тематична однорідність текстів», тому що, на думку авторів, саме ці характеристики мають найбільш практичне значення під час складання частотних словників. Текстові корпуси, необхідні для дослідження, формувалися на підставі наукових статей журналів відповідних спеціальностей США і Великої Британії, науково-технічної документації, газетних статей. Результати порівняльного аналізу семи текстових корпусів, що мають різні кількісні і якісні особливості, за критерієм однорідності/неоднорідності (його чотирма характеристиками) показали статистичні дані, які повною мірою відобразили індивідуальні особливості кожного з текстових корпусів.

Ключові слова: відрізок часу, область дискурсу, підмова, репрезентативність, текстовий корпус.

В статье рассматривается один из критериев лингвостатистики – критерий однородности/неоднородности, использование которого способствует созданию репрезентативных текстовых корпусов с последующей компиляцией статистически надежных частотных словарей (в статье таким словарем является частотный словарь «Радиоэлектроника»). Данный критерий проверялся путем сравнения семи частотных словарей относящихся как к научно-техническому дискурсу, так и к дискурсам других типов. Рассматривались следующие характеристики критерия однородности/неоднородности: «хронологическая однородность», «дискурсивная однородность», «однородность подъязыка», «тематическая однородность текстов», т.к., по мнению авторов, именно эти характеристики имеют наибольшее практическое значение при составлении частотных словарей. Текстовые корпусы, необходимые для исследования, формировались на основании научных статей журналов соответствующих специальностей США и Великобритании, научно-технической документации, газетных статей. Результаты сравнительного анализа семи текстовых корпусов, обладающих различными количественными и качественными особенностями, по критерию однородности/неоднородности (его четырем характеристикам) показали статистические данные, которые в полной мере отразили индивидуальные особенности каждого текстового корпуса.

Ключевые слова: область дискурса, репрезентативность, текстовый корпус, подъязык, отрезок времени.

Currently with the development of corpus linguistics compiling frequency dictionaries has become a simple and even a routine procedure, and their number and variety can satisfy any researcher.

However, such procedure seems to be easy and simple only at first sight. Since in forming the text corpus it is necessary to consider one of the most problematic aspects that can significantly affect the final results of the research – its (the text corpus') statistic reliability. This has been noted by many experts in linguistics [1–5].

First of all the problem of text corpus compilation should be touched upon. It requires to adhere to the statistical reliability principle which dictates the approximation of features of a limited set of texts, the researchers usually deal with, to the characteristics of the so-called general population. Regarded as belonging to such characteristics is the criterion (parameter) of homogeneity/inhomogeneity of the language as a whole, as well as of the types of discourse, types of domains of one and the same discourse, chronological homogeneity, etc.

In order to obtain a representative text corpus the scientists suggested to apply such statistical and qualitative methods as: 1) a continuous sampling of texts, which is considered more reliable to obtain authentic data than selective parts survey; 2) processing not one, but several fragments, i.e. corpora that relate to a particular or another genre, discourse or area of discourse, whereas one sample is too insignificant part of the general population; 3) comparison with the text corpora that are opposite in content or discourse [1–5].

Following the above recommendations for compilation of the most statistically reliable frequency dictionary the authors formed the one on the basis of text corpus "Radio-electronics" referring to scientific technical discourse with the method of continuous sampling of texts. The texts were selected from the scientific journals "Radio-electronics" published in the UK and USA. Other frequency dictionaries, the usage of which is provided by points 1 and 2 of the above methods, served as comparison objects for determining the statistical representation of the text corpus of the frequency dictionary "Radio-electronics".

The purpose of the paper is the following: to describe in terms of the homogeneity/inhomogeneity characteristic the comparative analysis of frequency dictionaries referred to various discourse types and various scientific discourse areas for determining the statistical reliability of the frequency dictionary "Radio-electronics"

In the presented paper the frequency dictionary created by the authors in the specialty "Radio-electronics" (it will be denoted as FDau in the paper) has been compared with the following frequency lists:

1) with two frequency dictionaries also formed on the basis of the scientific discourse texts and of the same knowledge domain as was chosen by the authors of the paper, i.e. 'Radio-electronics" (arbitrary notations FDub and FDal);

2) with two frequency dictionaries also belonging to the science and engineering discourse but describing other domains of science and engineering, namely, metallurgy (FDmet) and mathematics (FDmath);

3) with a frequency dictionary belonging to the scientific discourse as well but to the humanitarian science, that of psychology (FDps);

4) with a frequency dictionary formed on the basis of newspaper journalism texts, i.e. of the texts referring to another type of discourse (FDnews).

Though the homogeneity/inhomogeneity criterion includes eight (and in some papers even more than eight) characteristics, the framework of the article allows describing the comparative analysis of only some of them, which, in the authors' opinion, represent this statistical parameter in the practical experiment, concerning frequency dictionary compilation, most clearly and completely.

So, the characteristic "Chronological Homogeneity" of the homogeneity/inhomogeneity criterion is

considered first. As chronologically homogeneous one should acknowledge the texts related to some period of time of certain duration. The time length of such a span depends on the discourse type or the knowledge domain to which the analyzed text belongs. For instance, in fiction the period of time in which the productions can be viewed as chronologically homogeneous usually equals 30 years [6], as such a period seems socially significant and creatively active in the life activity of people pertaining to one and the same generation, i.e. born at the same time [7]. But such a regulation is relative. It means that the chronological limits of one generation involve the opportunity of cooperation of representatives of all contiguous generations, i.e. both those authors whose creative maturity falls on the vital activity of the given generation, and those whose creative period is connected with the past generation, as well as the authors whose creative activity falls on the new generation's life. Thereby the duration of a time period for chronologically homogeneous works of fiction can average no less than 25 and no more than 50 years.

To determine the duration of a time interval for chronologically homogeneous texts of scientific engineering discourse in this case it is usually associated with the differentiation of scientific discourse into the text sets of specialties (sublanguages) which reflect the studied branch of science or knowledge. If the case in point is a sublanguage that serves the established sphere of human activity, then the length of time for chronologically homogeneous texts can also be 30 years. However, if research is being performed on texts reflecting a new branch of science, especially the one connected not only with formation but with rapid development as well, then the length of time for chronologically homogeneous texts should be no more than five years.

The duration of a time span is still more reduced for the texts that can be considered chronologically homogeneous in the newspaper and journalistic discourse. The newspaper almost instantly registers all changes in the norm and usage due to the speech of the authors of newspaper texts who strive for maximum operability in communicating their materials reflecting the diverse and rapidly changing real life events.

They simultaneously seek to somewhat bring their use of language nearer to the reader's everyday speech, sometimes deliberately, sometimes unintentionally deviating from the norm and fixing new non-normative uses of linguistic units in speech, which reach fiction texts with a delay in years. Thus, the texts of newspaper and journalistic discourse are the source of the latest linguistic information.

It is this last feature of newspaper texts that makes it necessary to limit the chronological framework of the material being studied to one-two years.

To verify the reliability of the frequency dictionaries, with which the frequency dictionary "Radio-electronics compiled by the authors is compared by the homogeneity/inhomogeneity criterion, let us analyze the text corpora that served as the basis for creating these frequency dictionaries from the point of view of their (texts) publication. For the FDau dictionary, the material from journals over one year was used; for FDub – over three years, and for FDmet – over six years. As you can see, here we can talk about full compliance with the homogeneity/inhomogeneity criterion, since the two frequency vocabularies FDau and FDub, reflecting the rapidly developing scientific field as probabilistic-statistical models of Radio Electronics, have published periods of selected texts from one to three years. The frequency dictionary FDmet reflects a specialty that has long been traditionally included in the industry of any country, and is based on materials published over six years. The same can be said about the dictionaries FDmath, FDps where the duration of publication of articles does not exceed ten years. As for the FDal and FDnews dictionaries, their authors could not provide data on the period of publication of the texts underlying these dictionaries.

The following characteristic of the homogeneity/heterogeneity criterion, which was considered by the authors, is "Discursive homogeneity".

It should be noted here that in the present article the authors would not like to cover a sufficiently controversial questions about the nomenclature of discourses, styles and genres, but simply join the opinion that their classification includes the following types of discourses –fiction speech, newspaper and journalistic, scientific engineering (or just scientific), official and business discourse and, perhaps, the colloquial style.

Numerous studies in the field of discourse (discourse linguistics), describing both individual specialties and comparative analysis of several areas of discourse and performing statistical textual surveys, convincingly show that in each type of discourse the units of all levels of language structure – from phonemic to syntactic [9–18] – do not function in the same way, from which it can be concluded about the specific peculiarities of each type of discourse in terms of the functioning of language and the need to take into account the stylistic features in forming the linguistically homogeneous text corpus.

Evaluation of frequency dictionaries by the criterion of homogeneity/inhomogeneity usually relies on their belonging to the same type of discourse.

As it has already been mentioned, FDau, FDub, FDal, FDmet and FDmath are based on the texts of the technical areas of the scientific and technical discourse – electronics, metallurgy and mathematics; FDps also applies to scientific discourse, but to the humanities, psychology. Thus, six of the compared frequency lists refer to one and the same type of discourse – scientific one. This ensures the linguistic homogeneity of each individual dictionary, as well as of all the listed dictionaries among themselves.

As for the dictionary FDnews, it is based on the material of English newspapers and refers to the newspaper and journalistic discourse. Its (material) belonging to a discourse, other than the discourse of other dictionaries, does not allow it to be recognized as a homogeneous one in comparison with the texts of other six specialties.

The homogeneity/inhomogeneity criterion by the "Sublanguage Homogeneity" characteristics will be considered next.

The stylistic stratification of speech, reflected in the above types of discourse, is complicated by the identification of certain areas (sublanguages) or subsystems of the language, which are used under particular certain conditions of oral and written communication [19]. Each sublanguage reflecting a specific specialty should have its own speech unit statistics. Hence it is clear that when forming text corpora for linguistic-statistical (statistical linguistic) analysis it is also necessary to take into account their belonging to different sublanguages that serve different areas of knowledge, even those belonging to the same discourse.

Thus, with respect to the criterion under study, only the text corpora of the FDau frequency dictionaries (compiled by the authors), FDub and FDal can be considered homogeneous both within themselves and among themselves as all of them are compiled on the basis on radio electronics texts. The texts for the FDmet, FDmath and FDps dictionaries are homogeneous inside the discourse, but not among themselves, since they belong to different sublanguages.

Consequently, the entire set of text corpora for frequency dictionaries of scientific and technical discourse can be considered homogeneous from the point of view of the homogeneity criterion according to the characteristic "Sublanguage homogeneity" only within the common discourse. Between themselves, only FDau, FDub, and FDal corpora are homogeneous by this characteristic.

As for the texts for the FDnews dictionary, they are completely separate from all other dictionaries, and do not meet the criterion of homogeneity with any of the dictionaries in question.

The criterion of homogeneity/inhomogeneity by the characteristic "Thematic Homogeneity of Texts."

Among the linguistic features, which are specifically implemented and manifest differences between the texts, one should name a thematic feature, which reveals the relationship of the nature of speech with the topic, i.e. the contents of the text. The monothematicity of texts reflects their homogeneity according to the thematic criterion, and vice versa – the wider the scope of topics in the studied texts, the less homogeneous these texts will be.

It is believed that the need to adhere to one topic significantly reduces the lexical richness of texts, and that the dictionary compiled from texts reflecting many topics is richer than the dictionary of one author writing on one topic. However, if we want to get reliable results on the real number and nomenclature of words describing a given subject area, then it is necessary for the analysis to make a selection of linguistically homogeneous texts according to the thematic characteristic as well.

The analysis of frequency dictionaries performed by this characteristic showed that the text corpora for the FDau, FDal, FDmath, FDps and FDnews dictionaries are polythematic. This is due to the lack of thematic homogeneity in each of them, since all of them are based on texts, although belonging to the same sublanguage that serves one specialty, but their semantic space includes not one, but several sub-themes, which are integral parts of these specialties. Thus, the polythematic nature of the text samplings of these dictionaries determines their internal inhomogeneity.

The texts for the FDub frequency dictionary are homogeneous according to the analyzed criterion, since their distinctive feature is monothematicity, i.e. the material – scientific publications, technical documentation, materials of various firms from the US and UK scientific journals on electronics – was based on only one topic, and the semantic space of the text corpus was formed on the basis of only one topic.

The homogeneity/inhomogeneity criterion is quite difficult to apply to text corpora, the semantic space of which does not cover all, but only some of the sub-themes that are usually taken into account when creating the text set, which copies the semantic space as a percentage. An example of partial homogeneity or partial inhomogeneity is the FDmet frequency dictionary. It was analyzed as an object of study in order to determine the extent to which such partial fixation of the sub-themes of the semantic space can affect the further results of the statistical experiment.

Based on the above, we can draw the following conclusions.

1. For the formation of a representative text corpus, which will be the basis for a future statistically reliable frequency dictionary, all possible criteria must be applied, one of which is the homogeneity/ inhomogeneity criterion.

2. Although the criterion of homogeneity/ inhomogeneity of texts has many characteristics, however, according to the authors, those described in this article have the greatest practical importance in compiling frequency dictionaries: "Chronological homogeneity", "Discursive homogeneity", "Homogeneity of the sublanguage", "Thematic homogeneity of texts".

3. Frequency dictionaries, whose text corpora are presented for comparative analysis on the homogeneity inhomogeneity parameter of texts, have a variety of peculiarities in the homogeneity/ heterogeneity characteristics that were considered in the article. This situation may seem incorrect in the context of a comparative analysis of statistical data, where the units of comparison should be objects of the same nature. However, a comparison of exactly this type makes it possible to identify in practice which statistical characteristics are decisive for future results and which can be negligible. This was confirmed in the process of comparative analysis, in which text corpora with different quantitative and qualitative parameters showed different results.

In the future, it is planned to continue statistical studies of both text corpus units and the frequency dictionary "Radio electronics" units, as well as to draw for comparative analysis the dictionaries that appeared in the current study.

REFERENCES:

1. Piotrovsky Rajmund G. Quantitative Linguistics. An International Handbook / Rajmund G. Piotrovsky. 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). L. : Leningr. state. pedagogical. institute after A. I. Herzen, 1975. 120 p.

3. Захаров В.П. Корпусная лингвистика. Учебное пособие. 2005. URL : http://vp-zakharov.narod.ru/VictorZakharov_CorpLingv.doc.

4. Шубик С.А. Статистические методы в лингвистике. *Статистика речи и автоматический анализ текста.* Ленинград : Наука, 1980. С. 52–64.

5. Tuldava Ju. Исследования по сопоставительному и прикладному языкознанию. Tartu : Tartu Riiklik Ülikool, 1983. 164 p.

6. Статистичні параметри стилів. Київ : Наук. Думка, 1967. 260 с.

7. Шевырногова Л.А. Преимущественность поколений в поступательном развитии общества. Красноярск, 1983. 110 с.

9. Миляева Л.И. Структурно-семантическое исследование словообразовательных вариантов существительных в современном английском языке (парадигматический аспект) : автореф. дис. на соиск. науч. степени канд. филол. наук : спец. 10.02.04 «Германские языки». Пятигорск, 1984. 25 с.

10. Тарасова Е.М. Исследование употребительности словообразовательных суффиксов в научном стиле английского языка. Ленинград : Изд-во ЛГУ, 1972. С. 15–23. (Вопросы романо-германской филологии).

11. Бартко Н.В. Английские звукоизобразительные RL-глаголы: фоносемантический анализ : дис. … канд. филол. наук : 10.02.04. Санкт-Петербург, 2002. 288 с.

12. Дьяченко Г.Ф. Исследование семантики глагола в английских текстах подьязыков техники : автореф. дисс. на соискание научн. степени канд. филолог. наук : спец. 10.02.04 «Германские языки». Одесса, 1984. 16 с.

13. Лучак М.М. Вживання часових форм дієслова в сучасній англійській мові (на матеріалі трьох функціональних стилів) : автореф. дис... канд. філол. Наук : 10.02.04. Одеса : Одес. нац. ун-т ім. І.І. Мечникова, 2003. 20 с.

14. Неврева М.Н. Словообразовательная типология в подъязыках техники (на материале английского языка) : дисс. … канд. филол. наук : 10.02.04 «Германские языки». Одесса, 1986. 257 с.

15. Томасевич Н.П. Терминологическая лексика английского подъязыка автомобилестроения и ее взаимодействие с другими лексическими слоями : автореф. дисс. на соискание научн. степени канд. филолог. наук : спец. 10.02.04 «Германские языки». Одесса, 1984. 16 с.

16. Шапа Л.Н. Формы и функции имен прилагательных в научно-техническом тексте (на материале английского подъязыка электроснабжения) : дисс. … канд. филол. наук : 10.02.04. Одесса, 1989. 201 с.

17. Борисенко Т.И. Английские модальные глагольные конструкции в подъязыках техники : дис. … канд. филол. наук : спец. 10.02.04 «Германские языки». Одесса, 1989. 180 с.

18. Трофимова А.С. Синтаксические единицы современного английского языка в текстах делового стиля : дисс. … канд филол. наук : 10.02.04 «Германские языки». Одесса, 1986. 199 с.

19. Береснев С.Д. Субъязыки, стили и художественная литература. *Субъязыки и функциональные стили.* Ульяновск, 1980. С. 3–16.