

O. S. Meniailenko, SI "Luhansk Taras Shevchenko National University"

O. O. Blyskun, SI "Volodymyr Dahl East Ukrainian National University"

THE PROBLEM OF BUILDING A FOREIGN LANGUAGE LEARNING VOCABULARY

Meniailenko O. S., Blyskun O. O.

The Problem of Building a Foreign Language Learning Vocabulary

In this article, a method for building a learning vocabulary based on a probabilistic mathematical model of the words of the base minimum in the thematic material is developed. The method is as follows: form the words of the base minimum and carry out their frequency analysis; construct the density of probability distributions of the words of the basic minimum, on the basis of which determine the number of groups k ; divide the words into k groups by frequency and arrange lexicographically in each of the groups; groups of words of the basic minimum of a foreign language are in descending order. The proposed method of building a foreign language learning vocabulary reduces the time to study and memorize the words of a foreign language, as well as the probability of mistakes when translating into the user's language.

Key words: method, learning vocabulary, frequency analysis, mathematical model, mixture of distributions, ordering

1. Problem statement in general terms. Today, the main trends of modern society are the rapid development of information technology, the creation and active use of the Internet and social networks that are gaining popularity. All this activates the processes of globalization, unites people of different interests, causes the need for interpersonal and mass communication not only in real space, but also in virtual environments for "cyber relations" using different languages of the world.

However, foreign language training in educational institutions at different levels does not fully provide the required quality of training for pupils and students, which is confirmed by the actual practice of communication in foreign languages. Therefore, it can be argued that there is an objective problem of improving pedagogical systems and technologies of foreign language learning

related to the improvement of the quality of foreign language training of students, understanding and interpretation of foreign terms, etc., which makes relevant research in this area.

A special place in this problem is occupied by vocabulary dictionaries, the main purpose of which is to understand the didactic materials being learned and to remember words effectively. Today, a large number of approaches, methods for constructing various types of educational dictionaries, including online dictionaries, have been developed [1 – 10]. Their analysis is given below.

2. Analysis of the latest research and publications.

A known method for building foreign language learning vocabularies is used in the textbooks for translating texts and messages from a foreign language into the user's language, which is arranged in alphabetical (lexicographic) order for letters of a list of foreign language words containing lexical units provided with a translation into the user's language and comments. This method is implemented in numerous dictionaries, for example dictionaries by O.S. Akhmanova [1], V.K. Miuller and S.K. Boianus [2], etc.

Dictionaries built in this way have shown their effectiveness and versatility in learning foreign languages. However, their use in learning a foreign language, especially at an early stage, has serious difficulties due to the fact that they are not related to the topic or material being studied. This leads to increased time spent searching for and translating words from the foreign language being studied.

The method implemented in these dictionaries do not allow for the frequency characteristics of foreign language words and their peculiarities in the didactic material being studied, which complicates the memorization and comprehension of foreign language words in teaching material and also increases the likelihood of errors in translation. All this, in the end, increases the time to learn a foreign language.

Also it is known a method of building a foreign language vocabulary, which is used in the textbooks for learning a foreign language and translating texts from a foreign language into a user's language. Such a vocabulary consists of two parts.

The first part placing the initial samples of the text in the translation language divided into thematic sections, and in the second part, for each of the samples of text (thematic material), put in some way ordered, for example, alphabetically (lexicographically) words, turns and grammatical symbols ahmentiv translation, and, next to it, turns and grammatical fragments indicate the phonetic transcription graphical form words translated into the user's language and comments. This method is implemented in numerous dictionaries, for example, by R.R. Dickson [3], Ye.V. Nikoshkova [4] and others.

Vocabularies built in this way are one of the most common and effective languages for learning foreign languages. Tutorials using this method have been republished many times, for example, by R. Dickson's manual [3]. On the basis of this method is developing a large number of textbooks and dictionaries for the study of foreign languages and nowadays.

The main disadvantage of vocabularies that implement this method when learning a foreign language is that they do not take into account the frequency (probabilistic) characteristics of foreign dictionary words in the thematic sections (thematic material) to which they are associated. This makes it difficult to remember, understand, and reproduce the words of a foreign language in the subject matter being studied, as well as increase the likelihood of errors in translation, which ultimately increases the time to learn the thematic material and, consequently, the foreign language.

There is a method of "insitu" vocabulary building for translation into foreign language [5], which consists in compiling a vocabulary of two parts, with the first part placing the original samples of texts in the translation language, which are divided into thematic sections and marked with symbols of grammatical fragments of translation, and in the second part is placed alphabetically ordered words, turns and symbols of grammatical fragments of translation, and, next to the words, turns and symbols of grammatical fragments of translation indicate the coordinates of examples of the use of these words, turns and corresponding symbols of grammatical translation fragments in the texts of the first part of the dictionary.

The disadvantage of method [5] is the additional (redundant) information (page and line coordinates) that is used to relate to texts in the first part of the dictionary. Obviously, this information should not be remembered when learning words of a foreign language, however, when located next to the words of the dictionary, it is stored mechanically, through visual memory, in the memory of the learner of the foreign language (user). That is, in this case, it is a barrier to learning a foreign language, so it involves the irrational time spent filtering information, not to remember the coordinates of pages and lines.

Also, this method does not take into account the frequency (probabilistic) characteristics of foreign vocabulary words in the thematic sections to which they are related, which complicates the memorization, comprehension and reproduction of foreign language words and increases the likelihood of errors in translation. These disadvantages, in the end, increase the time to study the thematic section and the foreign language in general.

It is known method of building a vocabulary for translation from a foreign language [6] is that words of a foreign language selected from the vocabulary are placed on the storage medium sequentially, starting with the word beginning, before each word inscription, its semantic code and after each inscription is grouped by semantic codes of words that are similar in meaning, with the word being translated also having its semantic code, and possible translations of the word of a foreign language arranged in the sequence of semantic codes grouped according to their content.

The main disadvantages of this method [6] are: 1) the presence of additional information (semantic code), which is located next to the words of the vocabulary, and which should not be remembered in the process of learning a foreign language; 2) the frequency (probabilistic) characteristics of the vocabulary words in the thematic sections to which they are related are not taken into account; 3) not formalized (no description) procedures for determining the number of groups for the semantic codes in the vocabulary and the semantic codes themselves, which complicates the implementation of the vocabulary.

It is known a method of a vocabulary building of the alphabetic foreign language [7], with separate and / or labeled groups, which are combined sequentially applied, accompanied by translation and / or phonetic transcription of at least two words of a foreign language languages that contain sequentially spaced, end-to-end letters that are the same for all the words in the group, with all the words in the group containing the same number of the same letters.

The disadvantage of this method is the lack of direct linking of the vocabulary with the studied textual material, which does not allow to take into account the frequency characteristics of foreign language words, especially their use in the studied material. This makes it difficult to remember and understand the words of a foreign language, which ultimately increases the time to study it.

There is a method of a vocabulary building [8], which includes determining the base minimum of the words of the object language based on the frequency analysis of the use of common words, the distribution of the obtained basic minimum of the language into groups by parts of the language, the formation in these groups of smaller subgroups by thematic affiliation, in the received subgroups there are international words that have similarity in pronunciation and meaning with the words of the native language of the learner (student), which rely on in the initial stage of learning, and the received groups, etc. word groups are represented in the form of one or more tables on different types of data storage devices.

The disadvantage of this method of building a vocabulary of a foreign language [8] is the large number of groups and small subgroups into which words are broken, which significantly complicates the work with the dictionary, especially at the initial stage of learning the language, and therefore increases the time for learning it.

The disadvantage of the vocabulary building method [8] is that it is focused on the use of common words and does not generally involve a direct linking of the vocabulary to the material being studied (didactic section, topic, etc.).

Therefore, this does not allow for the frequency characteristics of foreign language words and their peculiarities in the studied thematic material, which complicates the memorization and comprehension of foreign language words and increases the likelihood of errors in translation. All of this ultimately increases the time to learn a foreign language.

Thus, the analysis of the basic approaches, methods of building foreign language vocabularies shows the relevance of research on the development of method of building educational vocabularies when learning foreign languages.

3. Selection of previously unresolved parts of the general problem. In the pedagogical problem of improvement of pedagogical systems and technologies of teaching in foreign languages the urgent task of developing a method of constructing educational dictionaries of a foreign language is solved.

The author's concept of the research is based on a probabilistic mathematical model of the words of the basic minimum in the thematic material, which is studied as a heteroscedastic mixture of their distributions [9; 11; 12].

This concept is confirmed by numerous statistical studies of foreign language vocabularies.

4. Formulation of the goals of the article (formulation of the task). The purpose of this work is to develop a method of constructing a vocabulary from a foreign language based on the probabilistic model of vocabulary words in the didactic material being studied.

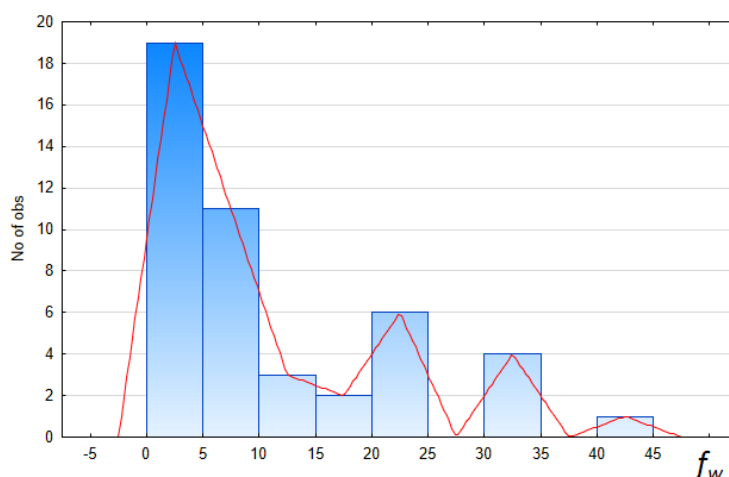
Research methods. Frequency analysis methods, classification methods, methods for building the probability density distribution of random variables were used to achieve this goal and to verify the reliability of the obtained results.

5. Outline of the main research material. The paper (research) deals with the problem of developing a method of building a foreign language vocabulary in which, due to the distribution of vocabulary words into groups and groups by frequency characteristics, which is performed on the basis of frequency analysis of the basic minimum of words of the thematic material of the language being studied. To solve this problem, a study of the frequency characteristics (density of

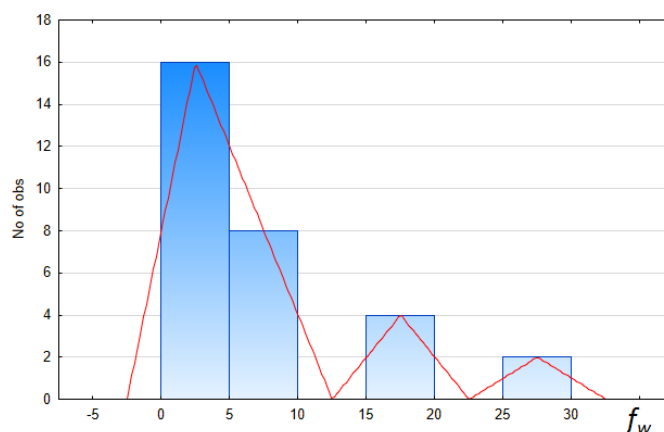
distribution) of words in the most common vocabularies in foreign language was conducted.

The process of constructing the density (polygon) of probability distributions based on experimental data is now generally accepted, it is widely reflected in various literature and publications, in addition there are a large number of application packages that allow it to be done, such as *Statistica* and others.

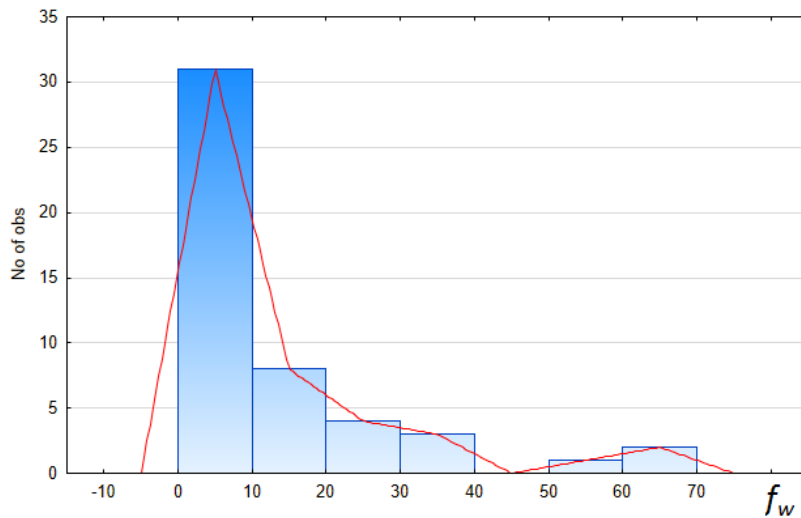
The Pictures 1–3 give examples of building the density (polygons, histograms) of probability distributions f_w of the words of the basic minimum (vocabularies) in the studied thematic material for the textbooks by R. R. Dixon [3] and Ye.V. Nikoshkova [4, p. 6], which were obtained through the Statistica 10 application package.



Pict. 1. An example of building the density (polygon) of the word frequency probability distribution f_w in the vocabulary for the *Lesson 2. The Family* [3, p. 40].



Pict. 2. An example of building the probability distribution of the frequency f_w words in the vocabulary for the *Lesson 3. A House and a Flat* [3, p. 50].



Pict. 3. An example of building the word frequency probability density distribution in the vocabulary for the *Lesson 1. Psychology as a science of conscious experience* [4, p. 6].

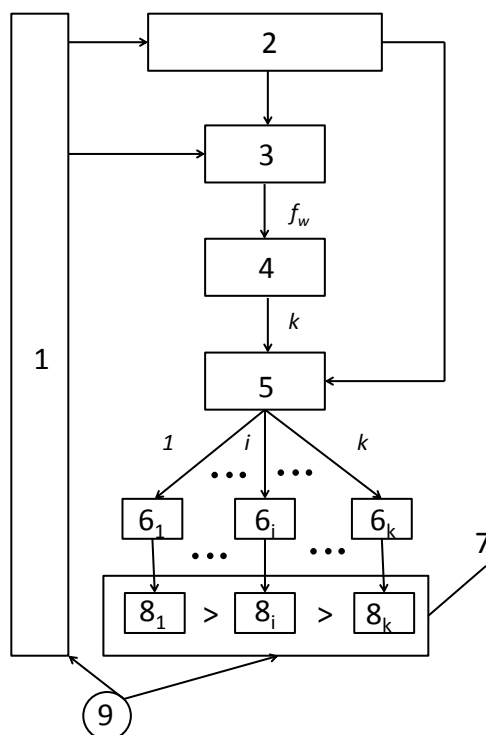
Theoretically, the resulting word frequency probability density distributions (see Pict. 1 – 3) represent a probabilistic mathematical model of the vocabulary. Such models are called heteroskedastic mixtures of probability distributions [9]. These mathematical models are well-known and quite fully reflected in various literature, for example in [9] etc.

The number of significant maxima in the heteroskedastic mixture of probability distributions allows us to determine the number k of statistically “independent” frequency groups that make up the vocabulary.

From the given examples (see Pict. 1 – 3) it follows that the word frequency probability density distributions f_w for the thematic materials (lessons) contained in Pict. 1 – four significant maxima, that is, the number of groups $k=4$, and for Pict. 2 and Pict. 3, respectively, $k=2$ and $k=3$.

Thus, the examples of frequency analysis (building the probability distribution density) for known educational dictionaries and thematic materials, shows the possibility of determining the number of groups in the dictionary, and also found that their number depends on both the features of the thematic material and the words of the base minimum (see Pict. 1 – 3).

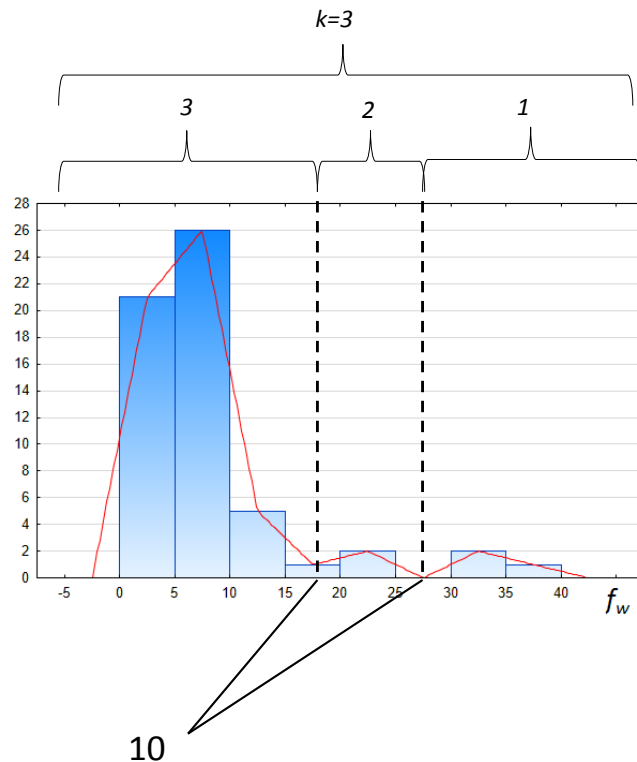
Based on the conducted research and on the probabilistic mathematical model of the vocabulary, a method of its building was developed, which is illustrated by the logical scheme of the sequence of actions, which is shown in the Pict. 4.



Pict. 4. The sequence of actions that implements the building a foreign language learning vocabulary, where: 1 - thematic material (educational text); 2 - base minimum words; 3 - frequency analysis of the words of the base minimum 2 in the thematic material and determination of the frequency f_w for each word of the base minimum; 4 - construction of the word frequency density distribution of the base minimum 2 in the thematic material 1 and determining the number of statistically "independent" frequency groups $1 \leq i \leq k$ that make up the vocabulary; 5 - division into statistically "independent" frequency groups of $1 \leq i \leq k$ words of base minimum 2; $6_1, \dots, 6_i, \dots, 6_k$ are words of the basic minimum of the foreign language divided into groups $1 \leq i \leq k$; $8_1, \dots, 8_i, \dots, 8_k$ - ordering lexicographically with phonetic transcription, translation into the user language and commentary of the base minimum foreign language words in each group ($6_1, \dots, 6_i, \dots, 6_k$); 7 - sorting the resulting groups $8_1, \dots, 8_i, \dots, 8_k$ in order of decreasing the word frequency in these groups and creating a dictionary; 9 - user (one who is learning a foreign language).

An example of determining the number of statistically "independent" groups k and the distribution of the words of the base minimum into groups $1 \leq i \leq k$ is shown in the Pict. 5. Here, vertical dividing lines 10 define the boundaries of groups that pass through the minimums of the probability distribution of words f_w

of the basic minimum of foreign language words (see Pict. 5, groups 1, 2 and 3 ($k=3$)). In addition to this method of word division into groups, there are various algorithmic methods that are not considered in this work [11; 12].



Pict. 5. An example of determining the number of statistically "independent" k groups and dividing the words of the base minimum by $1 \leq i \leq k$ groups by their frequency.

When $k=1$ (one group), the training vocabularies of the known manuals are obtained, that is, they are a separate case of the vocabularies built according to the proposed method.

The Pict. 6 shows an example of vocabulary 7 for the number of groups $k=3$, where statistically "independent" word groups are conventionally denoted by Arabic numerals 1, 2 and 3, and numbers 8_1 , 8_2 , and 8_3 denote words in groups that are lexicographically ordered. The user (one who is learning a foreign language), which is marked in the Pict. 5 by digit 9, begins the study of educational thematic material (see figure 1 in the Pict. 5) to identify and memorize the words of group 1,

which have the highest frequency of occurrence in the studied thematic material, sequentially moving to the words of groups 2 and 3.

WOCABULARY		
1		
do	[du:]	делать
see	[si:]	видеть
there	[ðeə]	там
2		
live	[liv]	жить
meal	[mi:l]	пища, еда
table	['teɪbl]	стол
3		
bath	[bɑ:θ]	ванна
divan	[di'væn]	диван
electric	[i'lektrɪk]	электрический
furniture	['fɜ:nɪʃə]	мебель
gas	[gæs]	газ
lamp	[læmp]	лампа
over	['əʊvə]	над
refrigerator	[rɪ'frɪdʒ(ə)reɪtə]	холодильник
set	[set]	приемник
wash	[wɒʃ]	умывать (ся)

Pict. 6. An example of a fragment of the vocabulary 7, constructed by the proposed method for the number of groups $k=3$, where the groups of words are conventionally denoted by the Arabic numerals 1, 2 and 3, and the numbers 8_1 , 8_2 , i 8_3 denoted words in the corresponding groups, which are lexicographically .

The proposed method (method) of building a a foreign language vocabulary has a cause-and-effect relationship, which is that the search, identification, study and memorization of words of the basic minimum of a foreign language begins consecutively, from the first group of words with their maximum frequency, to a group of words that have a minimum frequency of occurrence of words in the studied thematic material, thus, the words in groups are ordered lexicographically, which brings the dictionary closer to the thematic material, greatly simplifies the study and understanding of educational of thematic material.

Thus, foreign language vocabularies developed using the proposed method improve the memorization of foreign language words, reduce the time to learn them, and the likelihood of errors when translated into the user's language. In addition, the proposed method of building a foreign language learning vocabulary

can significantly simplify the use of the vocabulary, make it closer to the thematic material that is being studied, and therefore more effective when learning a foreign language.

6. Conclusions from the research and prospects for further research in this direction

1. Examples of frequency analysis of known educational vocabularies show that mathematical modeling of vocabulary word frequencies in thematic material are heteroskedastic mixtures of probability distributions that contain the number of k significant maxima (significant groups, classes) from one to four ($1 \leq k \leq 4$).

2. It is established that the number of k significant maxima (significant groups, classes) in the heteroscedastic mixture of probability distributions depends on both the features of the thematic material and the words of the basic minimum of the vocabulary.

3. A method of building a foreign language learning vocabulary is developed, which consists of the following: 1) form (distinguish) the words of the basic minimum for the thematic material being studied, in accordance with the educational goals and tasks; 2) conduct frequency analysis of the words of the basic minimum in the thematic material; 3) build the probability distribution of the word frequency probability of the base minimum in the thematic material; 4) determine the number of groups k by the number of significant maxima of the word frequency probability density distribution; 5) depending on the results of frequency analysis, the words of the base minimum are divided into k groups by frequency; 6) words of the basic minimum of the foreign language in each group are ordered lexicographically, provided with phonetic transcription, translation into the user's language and comments; 7) groups of words of the basic minimum of a foreign language are arranged in descending order of the frequency of words in these groups.

4. It is shown that with the number of significant maxima in the heteroscedastic mixture of probability distributions $k=1$, the vocabularies of known

textbooks are a special case of vocabularies built according to the developed method.

5. It is shown that the developed method of building a foreign language learning vocabulary can significantly simplify the use of the vocabulary, making it more approximate to the educational thematic material, and therefore more effective in the study of thematic sections (materials) from a foreign language. This improves the memorization of foreign language words, reduces the time for learning and memorizing them, and the likelihood of errors when translated into a user's language.

6. A promising area of research is the development of software tools for automated building a vocabulary in the developed way.

References

1. **Anglo-russkiy** slovar' [Tekst] : 20 000 slov / Pod red. Akhmanovoy O.S. – 22-ye izd. [English-Russian Dictionary [Text]: 20,000 words / Ed. by Akhmanova O.S. - 22nd ed.]. Moscow, *Soviet encyclopedia*. 1972. 614 p. (rus)

2. **Anglo-russkiy** slovar' [Tekst] = English-Russian dictionary: 40 000 slov / Sost.: V.K. Myuller, S.K. Boyanus [English-Russian Dictionary [Text]: 40,000 words / Comp. by V.K. Muller, S.K. Boianus]. K, *Kannon*. 1998. 1391 p. (rus)

3. **Dickson R.R.** Govorite po-angliyski. [Speak English]. Moscow, *Foreign Languages Publishing House*. 1961. 236 p. (rus)

4. **Nikoshkova Ye.V.** Angliyskiy yazyk dlya psikhologov: Ucheb. posobiye dlya stud. vyssh. ucheb. zavedeniy [English for Psychologists: Textbook for Students of Higher Educational Institutions]. Moscow, *VLADOS-PRESS Publishing House*, 2003. 160 p. (rus)

5. **Sposob** "tnsitu" postroyeniya slovarya dlya perevoda na inostranny yazyk: pat. 2054200 Rossiya: MPK G09B19/08. № 5048365/12; zayavl. 16.06.1992; opubl. 10.02.1996, Byul. №5. [Method "tnsitu" of building a dictionary for translation into a foreign language: US Pat. 2054200 Russia: IPC

G09B19 / 08. No. 5048365/12; pat. pend. on June 16, 1992; publ. on February 10, 1996, Bull. No. 5]. (rus)

6. **Sposib** pobudovy slovnyka dlya perekladu z inozemnoyi movy: pat. 60217A Ukrayina: MPK G 09B 19/08. № 2003032614; zayavl. 26.03.2003; opubl. 15.09.2003, Byul. №9 [Method of building a dictionary for translation from a foreign language: Pat. 60217A Ukraine: IPC G 09B 19/08. № 2003032614; pat. pend. on March 26, 2003; publ. on September 15, 2003, Bul. No 9.] (ukr)

7. **Slovyk** alfavitnoyi inozemnoyi movy: pat. 9031 Ukrayina: MPK G 09B 19/06. № 95125551; zayavl. 08.12.1995; opubl. 30.09.1996, Byul. №3. [Dictionary of the alphabetic foreign language: Pat. 9031 Ukraine: IPC G 09B 19/06. № 95125551; pat. pend. on December 08, 1995; publ. on September 30, 1996, Bul. No. 3]. (ukr)

8. **Sposob** obucheniya inostrannym yazykam: pat. 2150147 Rossiya: MPK G09B19/06. № 99112549/12; zayavl. 15.06.1999; opubl. 27.05.2000, Byul. №8 [The method of teaching foreign languages: US Pat. 2150147 Russia: IPC G09B19 / 06. No. 99112549/12; pat. pend. on June 15, 1999; publ. on May 27, 2000, Bull. No. 8]. (rus)

9. **Ivchenko G. I.**, Medvedev Yu. I. Smesi veroyatnostnykh raspredeleniy i sluchaynyye razmeshcheniya. Tr. po disk. matem. [Mixtures of Probability Distributions and Random Placements]. Moscow, 1998. T.2. Pp. 169 – 182. (rus)

10. **Reznikova O. O.** Rozrobka terminolohichnoho slovnyka anhlovnykh terminiv z psykholohiyi dlya ponyatiynoho aparatu ahresiyi v kiberprostorii [Development of Terminological Dictionary of English-Speaking Terms in Psychology for the Conceptual Apparatus of Aggression in Cyberspace]. Luhansk, *Publishing House of Luhansk Taras Shevchenko National University*. 2011. No. 12 (223), Part 2. Pp. 35 – 47. (ukr)

11. **Meniailenko O. S.** Suboptymal'ni adaptyvni alhorytmy identyfikatsiyi heteroskedastychnykh sumishey rozpodiliv [Suboptimal Adaptive Algorithms for Identification of Heteroskedastic Mixtures of Distributions]. Bulletin of Donetsk

University. Series A "Natural Sciences". 2006. Part 2, ed. 1. Donetsk, DNU. 2006. Pp. 399 – 405. (ukr)

12. **Sposib** rozpiznavannya obraziv: pat. 100078 Ukrayina: MPK G06 K 9/00. № 201414153; zayavl. 30.12.2014; opubl. 10.07.2015, Byul. №13 [Method for pattern recognition: US Pat. 100078 Ukraine: IPC G06 K 9/00. № 201414153; pat. pend. on December 30, 2014; publ. on July 10, 2015, Bul. No. 13]. (ukr)

Меняйленко О. С., Блискун О. О.

До проблеми побудови навчального словника іноземної мови

У роботі розроблено спосіб побудови навчального словника, що базується на ймовірнісній математичній моделі слів базисного мінімуму у тематичному матеріалі. Спосіб полягає у наступному: формують слова базисного мінімуму та проводять їх частотний аналіз; будують щільність розподілів ймовірності слів базисного мінімуму, на основі чого визначають кількість груп k ; поділяють слова на k груп по частоті та впорядковують лексикографічно у кожній з груп; групи слів базисного мінімуму іноземної мови розташовують у порядку убутання. Запропонований спосіб побудови навчального словника іноземної мови знижає час на вивчення і запам'ятовування слів іноземної мови, а також ймовірність помилок при перекладі на мову користувача.

Ключові слова: спосіб; навчальний словник; частотний аналіз; математична модель; суміші розподілів; упорядкування.

Меняйленко А. С., Блискун О. А.

К проблеме построения учебного словаря иностранного языка

В работе разработан способ построения учебного словаря, основанный на вероятностной математической модели слов базисного минимума в тематическом материале. Способ заключается в следующем: формируют слова базисного минимума и проводят их частотный анализ; строят плотность распределений вероятности слов базисного минимума, на основе чего определяют количество групп k ; разделяют слова на k групп по частоте и упорядочивают лексикографически в каждой из групп; группы слов базисного минимума иностранного языка располагают в порядке убывания. Предложенный способ построения учебного словаря иностранного языка в снижает время на изучение и запоминание слов иностранного языка, а также вероятность ошибок при переводе на язык пользователя.

Ключевые слова: способ; учебный словарь; частотный анализ; математическая модель; смеси распределений; упорядочения.

Information about the authors

Oleksandr Meniailenko – Doctor of Technical Sciences, Professor, Vice-Rector for Scientific and Pedagogical Work, State Institution “Luhansk Taras Shevchenko National University”. ORCID 0000-0001-6525-6849

Olena Blyskun – Candidate of Psychological Sciences, Assistant Professor, Department of Psychology, State Institution “Volodymyr Dahl East Ukrainian National University”. ORCID 0000-0001-7875-7714

The article was received by the Editorial Office on 01.03.2018.

The article was put into print on 30.03.2018.

Reviewer – V. V. Osadchy, Doctor of Pedagogical Sciences, Professor.