Automated approach to building the multilingual frequency dictionary on system analysis and computer technologies

I V Kovalev\textsuperscript{1,2,3,4,5}, M V Karaseva\textsuperscript{1,3} and A A Voroshilova\textsuperscript{5}

\textsuperscript{1}Siberian Federal University, 79, Svobodny Pr., Krasnoyarsk, 660041, Russia
\textsuperscript{2}China Aviation Industry General Aircraft Zhejiang Institute Co., Ltd, China
\textsuperscript{3}Reshetnev University, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russia
\textsuperscript{4}Krasnoyarsk State Agrarian University, 90, Mira Av., Krasnoyarsk, 660049, Russia
\textsuperscript{5}Krasnoyarsk Science and Technology City Hall of the Russian Union of Scientific and Engineering Association, 61, Uritskogo Str., Krasnoyarsk, 660049, Russia

E-mail: karaseva-margarita@rambler.ru

Abstract. The paper analyses the practical application of the multilingual adaptive – training technology. It contributes to the intensive accumulation of specialized foreign terminology by students who study one or more foreign languages for their professional purposes. The current version of the electronic multilingual dictionary is presented. The paper suggests a computer-aided approach to the analysis of the Multilingual frequency dictionary on system analysis and computer technologies.

1. Introduction
There exists a task to memorize large volumes of foreign terms in learning any foreign language. A vocabulary refers to terms, phrases, as well as grammatical forms presented in the form of phrases. The problem of foreign vocabulary is the most important for specialists working with scientific literature in a foreign language. Moreover, a scope of used terms is limited to the subject area chosen for study [1-5].

The preliminary processing of these texts should be carried out to point out the necessary vocabulary from texts of interest to a specialist, i.e., to compile a frequency dictionary of terms and phrases. The presence of such a dictionary allows taking into account the statistical laws of texts for learning vocabulary.

A frequency dictionary noted from other dictionaries for each output unit is accompanied by an indication of its frequency in texts used to compile this dictionary.

Frequency dictionaries can be classified depending on a language (languages), area of operation, input units.
First of all, a frequency dictionary is characterized by a vocabulary language it covers. There are dictionaries of Russian, English, German and other languages [6].

Multilingual frequency dictionaries take into account frequency properties of multilingual terms. As an example, a fragment of the multilingual dictionary on system analysis in electronic machinery, aviation and space industry [7] is presented.
One of the dropout strategies is to eliminate low-frequency terms. It means that a dictionary for speech perception is a frequency dictionary. For this reason, when the identity of a word relies mainly on formal, graphic coincidence, its semantics is not taken into account sufficiently. As a result, frequency characteristics are biased, distorted. That is why a frequency dictionary should be presented as clearly formalized lexical units, and well-established lexemes.

Traditional printed and even electronic dictionaries have a limitation as they are closed literary works and do not allow a user to modify any existing definitions, add new definitions or build new dictionaries.

A dictionary can be considered as an open one if a user can update easily the existing data in the dictionary, modify the entries stored in it, and also delete or add entries. A user can build a new dictionary, which will be filled with entries from another dictionary or new entries. One can add articles in other languages, including hieroglyphs. So, as a result, a new version of this dictionary, the Frequency dictionary on system analysis and compute technologies, with Chinese terminology necessary to collaborate with Chinese colleagues and teach Chinese students both professional vocabulary and the perception of lecture material and the implementation of practical and laboratory tasks [8].

Thus, the use of frequency dictionaries is the best alternative for learning foreign languages. These dictionaries help not only to circumvent such shortcomings of traditional electronic or printed dictionaries as the difficulties of finding derivative forms of the word and the closed nature of literary works, but also provide for the possibility of generating a selection of terms with the highest frequency [9].

### 2. Application of statistical methods in the analysis of specialized texts

Statistical methods penetrate wider and deeper into the most diverse fields of science, including those that were traditionally considered inaccurate, such as linguistics.

Information on the vocabulary application is provided by frequency dictionaries that register words, word forms or phrases found in the text (sample) examined for the dictionary compilation [10]. With these units, their frequencies are indicated in the dictionary, i.e. numbers showing how many times each dictionary unit has been encountered in this text.

To compile a frequency dictionary, it is necessary to have considerable period of time, familiarity with the statistical methodology of observations, certain linguistic qualification and knowledge of the language the studied text is written in.
The frequency dictionary building involves the application of linguistic and statistical analysis and it includes the following steps:

- definition of statistical elements (word, phrase, sentence);
- determination of the absolute frequency of elements in a single sample and in a total sample;
- calculation of the relative frequency and probability of the appearance of a language unit in the main set of terms of a professional sublanguage;
- checking the reliability of the obtained frequency characteristics by calculating standard deviations and relative errors;
- formalization of results in the form of lists, tables or graphs.

All words and phrases that make up the vocabulary of texts in a given specialty, as well as grammatical forms and syntactic constructions, have a certain probability of occurrence in these texts. It is also necessary to pay attention to their types in analyzing texts. The formation of the initial textual basis is based on the use of new general (non-special) journals, as well as reference books, essays, instructions for use and other types of texts [11,12].

The first result of statistical text processing is the absolute frequency. It shows how often the corresponding phenomenon occurs in the text being studied.

The relative frequency is a percentage that expresses a share of a language unit in the whole text. It is obtained by dividing the absolute frequency by the length of the sample. In other words, a relative frequency of a phenomenon is a ratio of the number of its actual occurrence to the number of its theoretically possible occurrence. If a sample is representative of the size of the subject language, then a relative frequency can be equated with the probability of a language phenomenon.

Linguistic statistics uses various lists, tables, graphs to display the results of the studies. Using a circular image and strip charts, parts are displayed as percentages. For graphical representation of quantitative features, such as the length of a word or sentence, histograms and a chain of polygons are suitable. Curves with a more or less smooth flow exceed this combination of frequencies in terms of quality and quantity. They make it possible to recognize functional dependencies between signs and their frequency.

Thus, frequency dictionaries make it possible to take into account the statistical properties of texts when forming more complete and high-quality information and terminological bases on their basis. It reduces significantly the time and human resources required to compile them, and also leads to faster and cheaper development of computer language learning systems [13,14].

3. Estimation of the multilingual frequency dictionary quality

In order to evaluate the quality of the obtained frequency dictionary, two parameters are usually determined. They are its reliability and efficiency [15,16].

The reliability is understood as the relationship between the relative frequency of the lexical unit (LU) in the dictionary and its probability in the general population, i.e., in the whole set of LUs used in this sublanguage. To this end, confidence intervals are calculated for the probabilities of LUs at an arbitrarily set significance level, as well as relative errors in determining the boundaries of these intervals. To calculate the boundaries of the confidence intervals of the probabilities of LU, the following formula is used:

$$P_{1,2} = \frac{F + \frac{1}{2}Z^2_\rho \pm Z_\rho \sqrt{F + \frac{1}{4}Z^2_\rho}}{N}$$  \hspace{1cm} (1)$$

where $P_{1,2}$ is lower and upper boundaries of the confidence interval; $\rho$ is a significance level constant; $F$ is absolute frequency of LU; $N$ is a sample size.

In mathematical statistics, a confidence interval of an estimated parameter is understood to mean such an interval with respect to which with a predetermined probability close to unity, called confidence
probability or at a given significance level, it can be argued that it contains an unknown value for this parameter. A confidence interval is random in both location and length.

If the experiment on building a frequency dictionary was set correctly, then in other potentially possible frequency dictionaries with samples completely similar to the given one; the relative frequencies of the same LUs with the same confidence probability will be within the calculated boundaries.

It should be taken into account that the exact form and volume of the confidence interval is determined by the distribution law of the studied random variable. In order to find the boundaries of this interval, it is necessary to identify a law of the frequency distribution of each LU in a large volume of texts. It requires a very much time and money.

The studies conducted to date on determination the nature of the frequency distributions of individual LUs according to sample texts have not revealed a universal law of the frequency distribution of the units examined. Moreover, it was found that the empirical frequency distributions of far from all LUs are subjects to the well-known theoretical laws of distribution of random variables. All this greatly complicates the task of finding reliable boundaries of confidence intervals of unknown probabilities of LUs [17].

The most common way out of mathematical statistics is to compute the boundaries of confidence intervals based on the assumption that a sample is normal. This approach is based on the asymptotic properties of estimation (in the limit, all known distributions converge to normal), and therefore it gives only approximate results. Moreover, it is suitable only for large volumes of samples and involves the use of a sample of finite sizes and obtaining instead of an unknown value of the parameter its estimated value, i.e., the frequency of the LUs.

Thus, the use of the confidence interval method for estimation the probabilities of linguistic quantities requires great care and deep insight into not only the phenomenon being studied, but also the mathematical model used, which is not always possible for a linguist without the help of a mathematician. The mathematician, in turn, is not always able to understand the language phenomenon well enough.

The extensive experience in statistical processing of experimental data and recording corrections to measured values without resorting to confidence probabilities has been accumulated in various fields of application of statistics.

When assessing the reliability of the frequency dictionary, one can also indicate a real error in the frequency of the LUs with which it was obtained, i.e., provided that the frequencies of the corresponding LUs are sufficiently large and homogeneous, in texts with a similar volume and content should not significantly go beyond the limits of the measured maximum error. The fact that confidence intervals are trustworthy can always be verified in practice, i.e., in new texts. For this purpose, it is enough to take several texts of the same length and subject line as the minimum samples used in composing frequency dictionaries with high-frequency lexical units. For low-frequency lexical units, such a check, unfortunately, will require processing of a long sample.

4. Conclusion

The frequency dictionary performance is usually evaluated empirically, i.e., as a degree to which a dictionary covers a random, similar text that is not included in the sample. Naturally, the longer a random text, the more reliable the results of such an assessment are. Due to the complexity of processing sufficiently long additional texts, in practice, they either compare equal in volume subsamples of the base sample, or take texts of a smaller volume.

Thus, to estimate the quality of a multilingual frequency dictionary, it is necessary to use not only methods of mathematical statistics, but also elements of linguistic and statistical analysis.

References

[1] Kovalev I V, Loginov Y Y and Zelenkov P V 2014 An Integrated System of Training Engineers for the Aerospace Industry in Siberia Using Innovative Technology of the Student Project
And-Team Work Proceedings Book of the 5-th International Conference of New Horizons in Education 2014 in Paris, France vol 1/5 pp 437–43

[2] Kovalev I V, Loginov Y Y and Zelenkov P V 2015 An integrated system of training engineers for aerospace industry in Siberia using innovative technology of the student project- and-team work Procedia - Social and Behavioral Sciences 174 537–43

[3] Kovalev I V, Loginov Y Y and Zelenkov P V 2015 Training of engineers in the aerospace university with application of technology research and education centers Turkish Onl Journ Educut Technol 14 717–23

[4] Kovalev I V, Loginov Y Y and Zelenkov P V 2015 The strategic approaches in quality of engineers training Turkish Onl Journ Educut Technol 14 561–5

[5] Kovalev I V, Loginov Y Y and Zelenkov P V 2016 Practice-Oriented Model of Engineering Education Turkish Onl Journ Educut Technol Special Issue for INTE pp 231–4

[6] Karaseva M V 1994 English-Russian frequency dictionary on system analysis SAA p 118

[7] Kovalev I V and Karaseva M V 2013 English-German-Russian frequency dictionary for system analysis in electronic engineering and aerospace industry SibSAU Krasnoyarsk p 216

[8] Kovalev I V and Karaseva M V 2020 Multilingual frequency dictionary on system analysis and information technology Siberian State University Krasnoyarsk p 124

[9] Kovalev I V, Piskorskaya S Yu, Karaseva M V and Voroshilova A A 2019 Computer-aided approach to synthesis of the frequency dictionary on system analysis in electronic machinery, aviation and space industry IOP Conference Series: Materials Science and Engineering 537(4) 042085

[10] Cromer V V 1997 Nuclear fan model of the vertical distribution of words in the Russian language Novosibirsk: NGPU No 52458

[11] Mueller Ch 1978 Initiation a la statistiquelinguistique Paris

[12] Kovalev I V, Kovalev D I, Karaseva M V, Pershakova K K and Tueva E V 2017 Multilingual environment of educational information interaction Scientific and technical information. Series 2: Information Processes and Systems 7 24–31

[13] Karaseva M V 2019 Computer-aided approach to synthesis the specialized frequency dictionaries Applied methods of statistical analysis. Statistical computation and simulation – AMSA‘2019 pp 421–7

[14] Engel E A, Kovalev I V and Engel N E 2016 Control of technical object on the basis of the multi-agent system with neuroevolution and student-teacher of-line learning IOP Conf. Series: Materials Science and Engineering 155 012001

[15] Kovalev I V and Loginov Y Y 2017 Opportunities of Interactive Teaching in the Implementation of Project Method Turkish Onl Journ Educut Technol 33 680–4

[16] Loginov Y Y, Kovalev I V 2017 Formation of research competence in university project-oriented training SHS Web of Conferences 37 1–7

[17] Kovalev I V, Loginov Y Y and Okuneva T G 2017 Education Quality Monitoring Of Students Of Technical And Economic Specialties European Proceedings of Social & Behavioural Sciences (EpSBS) XXXI 579–88