IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

Computer-aided approach to synthesis of the frequency dictionary on system analysis in electronic machinery, aviation and space industry

To cite this article: I V Kovalev *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **537** 042085

Computer-aided approach to synthesis of the frequency dictionary on system analysis in electronic machinery, aviation and space industry

I V Kovalev[1,2,3], S Yu Piskorskaya[1], M V Karaseva[1, 2] and A A Voroshilova[1,3]

[1]Reshetnev Siberian State Aerospace University 31, KrasnoyarskyRabochy Av., Krasnoyarsk, Russian Federation
[2]Siberian State University, 79, Svobodny Av., Krasnoyrsk, Russian Federation
[3] Krasnoyarsk Science and technology city hall, 61, Uritskogo ave., Krasnoyarsk, 660049, Russian Federation

E-mail: karaseva-margarita@rambler.ru

Abstract. The practical use of the multilingual adaptive - training technology contributes to the intensive accumulation of specialized foreign terminology by students who study one or more foreign languages for their professional purposes. Electronic frequency dictionaries, built on a multilingual principle, are the basic components for supporting the multilingual adaptive – training technology. The article considers the computer-aided approach to the analysis of Frequency Dictionary on System Analysis in the Electronic Machinery, Aviation and Space Industry.

1. Introduction

The purpose of the information multilingual adaptive - training technology is the intensive accumulation of specialized foreign terminology of students and specialists studying one or more foreign languages for their professional purposes. The main components of its support tools are electronic frequency dictionaries, built on a multilingual principle, and computer systems that implement the algorithm for training terminological vocabulary [1-5].

A frequency dictionary (information basis) is the information, whose portions are given to the student. The student will learn quicker and better the words and phrases that are found in the texts more often and these words are from narrow subject field. That is why it is necessary to take into account the frequency properties of texts. Table 1 shows an example of the information presented in the frequency dictionary [6]:

Table 1. Portion of the frequency dictionary.

Eng	Ru
Decompositionalgorithm, 49	Алгоритм разложения, блочный алгоритм, 15
Correlation analysis, 170	Корреляционный анализ, 110

Multilingual frequency dictionaries take into account the frequency properties of multilingual terminology [7]. An example of a portion of a multilingual dictionary on system analysis can be observed in table 2:

Table 2. Portion of the multilingual frequency dictionary.

Eng	De	Ru	Fr
ability, 4	Fähigkeit, 4 f	способность, 8	Lacapacité, 5
above, 32	über, 100	выше, свыше, 19	Plushaut, 44
abstraction, 7	Abstraktion, 1 f	абстракция, 42	L'abstraction, 15

The dictionary in a broad sense is the most essential component of the speech perception model.A main operational unit is a word while speech perceiving. It follows that, in particular, every word of the perceived text should be identified with the corresponding unit of the student's (or a reader) internalvocabulary. It is natural to assume that from the very beginning the search is limited to some sub-fields of the dictionary. In a typical case the actual phonetic analysis of the sounding text gives only some partial information about the possible phonological appearance of the word, and this kind of information is obliged by a certain set of terminology; therefore, the problem arises (a) to select the corresponding set according to one or another parameter and (b) within the determined set (if it is adequately selected) to produce a “sifting out” of all words, except for the only one that corresponds best to the given word of the recognized text. One of the "dropout" strategies is the elimination of low-frequency words. It follows that the terminology for speech perception is a frequency dictionary. Partly for this reason, when the word identity is largely based on formal, graphical coincidence, semantics is taken into account insufficiently. In the result,frequency characteristics are biasedand distorted. For example, if the words from the combination of "each other" areincluded by a compiler of the frequency dictionary in the general statistics of the use of the word "friend", then it is hardly to be justified. We are torecognize that in the word-combination we already have other words, and it is more precise that a separate vocabulary unit is a whole combination taking into account the semantics.That is why training is provided for both clearly formalized lexical units and well-established lexemes [8-10].

When organizing electronic frequency dictionaries that are the information base for supporting the multilingual adaptive-learning technology, the following objectives are pursued [10]:

- to reflect some of the important qualitative and quantitative aspects of general terminology on system analysis in English, German, French and Russian, resulting from statistical text analysis and description;
- to promote the organization of vocabulary learning and vocabulary accumulation in a rational way with the help of computer interactive training tools.

2. Texts analysis for the formation of electronic frequency dictionaries

The emerging directions of linguistics are connected with some applied tasks. At present, the direction of the applied linguistics has been practically formed. It is connected with the foreign languages training. Twenty years ago, a language description in teaching was considered as a problem of a methodology. Now it is clear that a methodology has other tasks, and a description of language phenomena should be carried out within the framework of linguistics and according to its laws. It is necessary to pay attention to the fact that the situation is similar to the way the direction of “computational linguistics” developed. Initially, problems connected with the automatic training were entirely attributed to the competence of programmers, and only then, the applied development of machine translation, automatic referencing, etc., was recognized as a linguistic value [11–13].

In the course of the language description some independent results were obtained. And, moreover, some developments that had appeared earlier, for example, functional grammar, linguistics of a coherent text, and possibly others, fell into the mainstream of this new direction.

Consider the peculiarities of pedagogical linguistics in connection with the problems facing this field.

The main peculiarity of the language description for the training purposes is to take into account the psychological properties of a person related to speech activity. These peculiarities are associated with the properties of memory (memorization, storage of information, its activation), understanding of speech and its generation and with peculiarities of communication, considered in a broader aspect (social, interpersonal relations, etc.). If we compare this psychological information with information about the structure of the language represented by the grammar, in many cases inconsistencies will be found. This fact is known to all teachers of a non-native language (especially if it is native to the teacher). First of all, they concern active speech activity since it requires the use of information on the compliance of language structures with the intentions of the speaker [9].

Language statistics can be defined as an auxiliary discipline of linguistics that explores the quantitative aspects of the language system use, including a professional-oriented system. Previously, some scientists used statistical methods successfully [14,15]. Language statistics supplements the qualitative methods of language description through additional data characterizing the frequency of language phenomena. It is very useful in such practical field as information retrieval, lessons in foreign languages. Mathematically, this approach helps to model professional-oriented language communication as a probabilistic process, allowing one to determine the objective parameters of language differentiation expressed in various sublanguages, professional languages, professional dialects or styles.

Linguistics uses statistical methods, primarily in problems covering the language functionally, in texts, gathering separate passages into a coherent text. It is absolutely impossible to do in any other way because of the wide variety of language communication in various professional fields.

When automating a general statistical analysis, the following steps can be distinguished: definition of statistical elements (word, phrase, sentence); determination of the absolute frequency of elements for a single sample and total sample; calculation of the relative frequency and probability for the main body of professional vocabulary terms; validation of the obtained frequency characteristics by calculating the standard deviations and the relative error; formalization of results in the form of lists, tables or graphs; interpretation and synthesis of results, up to the formulation of patterns.

Since it is almost impossible to cover the whole unity of subject-language communication even for only one language and one area, subject-language statistics should be based on the most representative sample tests, i.e. on the written or oral typical special texts. Each linguistic-statistical analysis begins with the selection and preparation of the corresponding text base. When specific tasks are set in the framework of applied linguistics, for example, when defining terminology for learning in a foreign language lesson or when compiling vocabulary for internal documentation, the text base can be very limited.

It is also necessary to pay attention to the type of the texts. Textbooks of higher and vocational schools of a general character are particularly suitable for the definition of a scientific and technical basic terminology. They guarantee a systematic, proportional and complete coverage of the material and the necessary language tools for its presentation. Moreover, they are less influenced by the individual language use by individual representatives of the profession. The further formation of the text base is built using new journals of a non-special nature. Reference books, reports, progressive messages, and instructions for application other types of texts, by contrast, are a favorable starting point for observing subject-language peculiarities at the level of both a sentence and a text.

The first result of statistical text processing is absolute frequency. It shows how often the corresponding phenomenon occurs in the text under study. However, it has little value for further research in the practical use of the results or in general for generalized statements, since it directly depends on the size of the selected text. It serves mainly as an initial value, for example, for calculating the relative frequency.

The relative frequency is a percentage that expresses the proportion of a language unit in the whole text. It is obtained by dividing the absolute frequency by the length of the sample, for example, for a word with a particular 186 in one sample from N = 50,000, the relative frequency will be calculated as 186 / 50,000 = 0.00372.

In other words, the relative frequency of a phenomenon is a ratio of the number of its actual occurrence to the number of its theoretically possible occurrence. It is possible to equate the relative frequency to the probability of a linguistic phenomenon if a sample is representative for a subject language. Then it gives grounds for statements about the statistical structure of the relevant sublanguage or about the importance of individual elements for the text organization.

A particularly important step in the linguistic and stylistic analysis is the reliability control of the determined data. There exist various ways of control. The standard deviations (errors), the relative error, and the confidence interval are primarily taken into account in stylistic statistics and subject-language statistics.

Standard error (mean square error) is a measure of the variability of the average frequency of a linguistic phenomenon in partial sampling. Its calculation is done according to the formula:

$$S = \sqrt{\frac{SAQ}{n-1}}, \tag{1}$$

where *S* is a standard error; *SAQ* is a sum of the square of the error; *n* is a number of control samples.

The relative error is calculated for certain lexical units in frequency dictionaries to determine the accuracy of these dictionaries. It is done according to the formula:

$$|f-p| = Zp\sqrt{\frac{p(1-p)}{n}}, \tag{2}$$

where *f* is a relative frequency; *p* is a probability; *Zp* is a coefficient for a given level of confidence *p*; *n* is a volume of the sample control.

In papers concerning with the linguistic, the simplified versions of this formula are used. They assume that the difference $(1-p)$ is approximately equal to one for a small p. A general variant of determining the relative error is as follows:

$$\delta = \frac{Zp}{\sqrt{nf}} \text{ or } \delta = \frac{Zp}{\sqrt{F}}, \tag{3}$$

where δ is a relative error; *Zp* is a coefficient for the given level of confidence *p*; *n* is a sample size; *f* is relative frequency; *F* is absolute frequency.

The calculation of the confidence interval is a sophisticated version of the calculation of the relative error. The lower and upper limits ($p_1$and p_2) of the oscillations and the average frequency are determined. There exist different methods of calculation, for example:

$$p_1 = \frac{fN + \frac{1}{2}Zp^2 - Zp\sqrt{f(1-f)N + \frac{1}{4}Zp^2}}{n + Zp^2}, \tag{4}$$

$$p_2 = \frac{fN + \frac{1}{2}Zp^2 + Zp\sqrt{f(1-f)n + \frac{1}{4}Zp^2}}{n + Zp^2}, \tag{5}$$

χ^2-test, determines whether any observed differences occur in different samples, or whether the samples belong to the same basic population (functional style, sublingual, objective language, type of text, etc.). In most cases, this involves the verification (authenticating or falsifying) of the main assumption (null hypothesis); for example, the expectation that word types play a similar role in the text. The reference value$\chi$is a sum of the observed and expected frequencies for a certain number of variables referred to the expected frequencies

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{bi} - f_{ei})^2}{f_{ei}}, \tag{6}$$

where k is a number of variables; i is a variable; $f_{bi}$is the observed frequency of the variable; f_{ei} is the expected frequency of the variable.

When comparing samples, the expected frequency of f_{ei} is usually equated to the average frequency$\overline{x}$.

$$\chi^2 = \frac{\sum(x_i - \bar{x})^2}{\bar{x}}, \tag{7}$$

The subject-language statistics use various lists, tables, and graphs to present the research results. With the help of a circular image and strip charts first of all, parts in percentage values are given. Histograms and a chain of polygons are suitable for the graphic representation of quantitative peculiarities, such as word length or sentences length. Curves with a relatively typical flow exceed this simple combination of frequencies in qualitative and quantitative terms. They help to recognize the functional relationships between signs and their frequency. The frequency of language phenomena can itself become a sign characterized by other data.

There exist, for example, the following dependencies:

- between the frequencies of lexical units and their classes in the frequency dictionary;
- between the frequency and its probability of occurrence in the text;
- between the frequency and a relative error;
- between classes of one frequency dictionary and the cumulative number of lexical units;
- between classes and class wrappers;
- between the frequency and potency of communication;
- between frequency and degree of specialization of the subject lexical vocabulary;
- between the length of the text and the amount of vocabulary, etc.

Linguo statistics helps to determine which language phenomena occur in speech or texts more often. Statistical methods study the vocabulary of the language intensively. Frequency dictionaries givethe information about the general vocabulary. A frequency dictionary registers words, word forms or word-combinations that have been encountered in the text (sample) studied for its compilation. Together with these units (i.e., words, word forms, word combinations), their frequencies are indicated in the dictionary, i.e. numbers show how many times each dictionary item has been encountered in the given text [6, 7].

The compilation of any frequency dictionary requires some time. A developer is to be familiarity with the statistical method of observation. For example, developers of the frequency English-Russian dictionary-minimum of newspaper vocabulary [9] applied the following methodology. The dictionary was compiled on the basis of linguistic statistical analysis of the language of newspapers and magazines in the UK and the USA. Texts were selected with a total length of 200,000 words from different newspapers and magazines. Words and word-combinations with varying degrees of stability were manually written out of these texts. Nowadays, the compilation of frequency dictionaries is realized with the help of a computer.

A text has a statistical structure. Its essence is that all the words and word-combinations 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 the texts of this special area [11].

On the other hand, if a special text is broken up into small portions, then one part of the linguistic units will give approximately the same frequencies in these texts, thus revealing the stability and uniformity of use. Another group of linguistic units gives unstable and uneven use in separate text portions. The first group is usually composed of auxiliary words and commonly used word-combinations. The second group is mainly formed by words and word-combinations directly related to the content of the text of the given special area (these words and word-combinations are often called keywords).

The texts structure of different special areas is not the same. The probabilities and distributions of keywords and word-combinations are noticeably different, while the statistics of auxiliary words and some commonly used words and word-combinations remain unchanged.

It is possible to find regularities in vocabulary functioning of the given language and get an idea of its quantitative structure if one studies the text of the sufficiently large volume. Such an analysis reveals, for example, two important linguistic-statistical regularities [16].

1. In any text, no matter how large it is, only a small portion of the vocabulary is used. Obviously, the child's vocabulary is much poorer than an adult's vocabulary. But the results of the experiment are surprising. Only 25,000 different word forms turned out to be in 100,000 letters, classroom and homework with a total length of 6 million words used (data from the analysis of written speech of schoolchildren).

Special scientific, technical and journalistic texts are also quite different in terms of vocabulary volume. The analysis of texts in English, Romanian and Moldavian languages showed that the vocabulary of nonfiction texts is approximately 2.5 times larger than the vocabulary of special texts. These numbers indicate that in different areas of speech communication, different numbers of words are used.

2. The second linguistic regularity is that even a limited part of the vocabulary of the language is used irregularly in the speech (text). Some words are used more often, others are used more often less often, and most of the text is an insignificant number of the most frequent words. For example, the following results were obtained when recording and analyzing telephone conversations: 737 most frequent words are over 95% of all the word used [10].

3. Conclusion

As it has been already noted, the frequency dictionary indicates a number of cases of the word use in texts that were analyzed to compile a dictionary. Frequency dictionaries differ depending on the principle of position of the database position. Words or word-combinations can be placed alphabetically, as in a typical dictionary, with its frequency to a word. Also, words and word-combinations can be arranged in descending order of frequencies, starting from the most commonly used word. The first version of the dictionary is addressed to a student; the second one is to the learner. A student can also work with the second version of the dictionary to learn a foreign language independently, for example, when memorizing words and word-combinations in portions, depending on their frequency or checking proficiency in terminology units, starting with the most frequent ones.

References

[1] Kovalev I V, Loginov Y Y amd 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 Paris France* **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 - Socialand Behavioral Sciences* pp 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. Educat. 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. Educat. Technol* **14** 561–5

[5] Kovalev I V, Loginov Y Y and Zelenkov P V 2016 Practice-Oriented Model of Engineering Education *TOJET The Turkish Online Journal of Educational Technology* pp 231–4

[6] Karaseva M V 1994 *English-Russian frequency dictionary on system analysis (*Krasnoyarsk: SAA) p 105

[7] Kovalev I V and Karaseva M V 2013 *English-German-Russian frequency dictionary on systems analysis in electronic engineering and aerospace* (SibSAUKrasnoyarsk) p 216

[8] Kromer V V 1997 *Nuclear Fan Model of Vertical Word Distribution in Russian* (Novosibirsk: INION RASNSPU) pp 132–46

[9] Mueller Ch 1968 *Initiation a la statistiquelinguistique* (Larousse) p 249

[10] Kovalev I V, Kovalev D I, Karaseva M V, Pershakova K K and Tueva E V 2017 Multilingual environment of information and educational interaction *Scientific and technical information Series 2: Information processes and systems* **7** 24–31

[11] Engel E A, Kovalev I V and Engel N E 2016 *IOP Conf. Ser.: Mater. Sci. Eng.* **155** 012001

[12] Kovalev I V, Losev V V, Saramud M V, Kuznetsov P A and Petrosyan M O 2017 To the Question of the Organization of a Learning Environment for Developers of Cross-Platform OnBoard Software for Unmanned Aerial Vehicles *The Turkish Online Journal of Educational Technology* pp 700–5

[13] Kovalev I V and Loginov Y Y 2017 Opportunities of Interactive Teaching in the Implementation of Project Method *The Turkish Online Journal of Educational Technology* pp 680-4

[14] Loginov Y Y and Kovalev I V 2017 Formation of research competence in university project-oriented training *SHS Web of Conferences* **37** 01027

[15] Kovalev I V, Loginov Y Y and Okuneva TG 2017 Education Quality Monitoring Of Students Of Technical And Economic Specialties *European Proceedings of Social &Behavioural Sciences EpSBS* pp 579–88

[16] Alekseev P M 1971 *Frequency English-Russian dictionary-minimum on electronics* (Moscow: Voenizdat) p 300