Multi-linguistic analysis of the uniqueness of scientific and educational texts based on lexically related components

I V Kovalev1,2,3,4, D I Kovalev2,5, A A Voroshilova3, E V Tueva1 and E V Tuev3

1 Siberian Federal University, 79, Svobodny pr., Krasnoyarsk, 660041, Russia
2 Krasnoyarsk State Agrarian University, 90, Mira pr., Krasnoyarsk, 660049, Russia
3 Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russia
4 China Aviation Industry General Aircraft Zhejiang Institute Co., Ltd, Huzhou, Zhejiang Province, China
5 Krasnoyarsk Science and Technology City Hall, 61, Uritskogo Str., Krasnoyarsk, 660049, Russia

E-mail: kovalev.fsu@mail.ru

Abstract. The article presents computer modeling and the development of algorithmic and information tools to support multilingual analysis of the uniqueness of scientific and educational texts using the technique of lexically related components. This new educational and communication methodology allows the formation of interlanguage and intralingual associative links. The authors carried out a theoretical and informational analysis of the communication structure and information base in computer modeling of the process of multilingual analysis of the uniqueness of texts. The development of algorithmic and information tools for supporting multilingual analysis of the uniqueness of texts on the basis of lexically related components with the formation of interlanguage and intra-lingual associative links has been carried out. As a result of the work, a new method of computer modeling of multilingual analysis of the uniqueness of texts is proposed, focused on the formation of intralingual and interlanguage associative links. To implement the proposed methodology in the process of multilingual analysis of the uniqueness of texts, the structure of frequency lexically related dictionaries has been developed as a means of information and terminological support.

1. Introduction
The development of modern science and technology, including production technology, information technology, invariably entails the emergence of new scientific texts of a descriptive nature [1-4]. Scientific articles, monographs, technical documentation, textbooks, self-study guides are far from the whole range of scientific, educational literature and Internet sources accompanying the development of new scientific directions and new learning technologies, software products, as well as certain changes in information and technological processes [5-7].

The huge volumes and intensity of the emergence of new scientific and special texts, as a rule, in different languages (depending on the country of origin, the place of residence of the authors, etc.) and the need for communication with foreign partners, both in the production and scientific spheres, often force analyze the text for uniqueness (in recent years, the requirements for correct authorship and
uniqueness of texts have been tightened both by the academic community and by Internet users and manufacturers of Internet resources, in almost all areas of knowledge) [8-10].

As part of solving this problem, when translating a text into foreign languages, in many cases, the professional orientation of the text should be taken into account. One of the main points in this case is the use of a specific scientific / educational specialty, research direction and a number of new technologies [11].

The modern development of computer modeling and information technology today makes it possible to develop an effective software and algorithmic apparatus for creating computer interactive systems for multilingual analysis of the uniqueness of texts, including when using multilingual foreign terminological vocabulary [12-14]. At the same time, a modern approach is used within the framework of multilingual technology, based on lexically related components both within one linguistic field and at the level of interlanguage interaction.

The multilingual technology (ML-technology) considered by the authors was developed specifically for the creation of such systems. It is based on the adaptive model of L.A. Rastrigin [15] and electronic multilingual frequency dictionaries. Its advantages, from the point of view of analyzing the uniqueness of texts, include the ability to generate associative fields between analogous terms of different languages [16]. However, ML-technology, like many others created for teaching [17-22], does not fully use lexical dependencies within one or many languages. Therefore, the generation of intralingual associative fields occurs exclusively at the stage of applying knowledge in practice and has a spontaneous character. If such generation takes place in an organized manner and directly in the real-time process, the quality of such a process, and, consequently, of computer systems, will increase significantly.

The described problems in the field of science and education, in information technologies determined the relevance of the work, which consists in the need for computer modeling and construction of a methodology for the process of multilingual analysis of the uniqueness of texts on the basis of lexically related components. This makes it possible to develop an appropriate model-algorithmic apparatus and information support, which gives scope for further research.

2. Computer model of the text analysis process based on lexically related components

2.1 Electronic frequency subject-oriented dictionaries

The purpose of using computer modeling of the process of multilingual analysis of the uniqueness of texts is to identify the identity of text fragments of specialized, including foreign text, both in one and in several foreign languages. This can be done for educational, scientific and other professional purposes. The main components of the tools for supporting this process are electronic frequency dictionaries (EFD), built on the multilingual principle, and computer systems that analyze the uniqueness of texts based on specialized terminological vocabulary [23-29].

EFD is exactly that information, portions of which are used to establish the identity of text fragments, and the frequencies allow you to find the most common analogs in other languages, that is, those words and phrases that are most often found in the texts of a specialized subject area. This allows, when checking the uniqueness of a text in one language, to identify analogues in another language, and taking into account the frequency properties of the texts of this language, to draw a conclusion about the degree of uniqueness of the text in the event that it is translated into this language.

Table 1 shows an example of the information representation in the frequency dictionary.

| Table 1. Fragment of the frequency dictionary. |
|-----------------------------------------------|
| combined method, 49 | kombinirovannyy metod |
| research, 1020 | issledovaniye |
| mathematical analysis, 177 | matematicheskiy analiz |
| expert approach, 110 | ekspertnyy podkhod |
Multilingual frequency dictionaries take into account the frequency properties of multilingual terms [18]. An example of a fragment of the multilingual dictionary for system analysis can be seen in table 2.

| tool, 61 | Werkzeug, 4 n, Mittel, 17 n | sredstvo, 100, instrument, 1 |
| basic tool, 61 | Hauptmittel, 17 n | osnovnoye sredstvo, 100 |
| topic, 9 | Thema, 4 n, Frage, 23 f | tema, 9, vopros, 62 |
| topology, 13 | Topologie, 1 f | topologiya, 8 |
| total, 129 | Ergebnis, 13 n, zusammenfassend, 2 | itog, 44, polnyy, 190 |
| tracing, 9 | Nachführung, 1 f | fslzheniye, 1 |

The dictionary is an essential component of the model for establishing the identity of texts in different languages [30-39]. When analyzing texts, the word acts as the main operational unit. It follows from this, in particular, that each word of the analyzed text must be identified with the corresponding vocabulary unit of the internal dictionary of this text. It is natural to assume that already at the initial stage, the search for words when comparing them is limited to some subdomains of the dictionary.

When organizing EFDs, which constitute an information base for supporting multilingual technology, the following goal is pursued: to reflect some important qualitative and quantitative aspects of common vocabulary in various terminological areas (for example, in systems analysis for the aerospace engineering in English, German, Russian and Chinese, obtained as a result statistical analysis and description of texts [40]).

As a result of the conducted statistical research and description of texts on system analysis for aerospace engineering of about 30000 words, a list of words of more than 2500 was compiled, the adequacy of which was established for all four languages, and then about 2000 words were included in the frequency dictionary, which is a component of a computer system implemented in based on ML-technology. Since almost all frequency dictionaries are monolingual, which significantly reduces the degree of word activity at a certain word meaning in the dictionary minimum, an analysis of multilingual texts was carried out at a systematic level. The dictionary is organized as English-German-Russian (and Chinese in the latest edition of 2020), however, the principle of its construction in the form of a multilingual information and terminological base (ITB) makes it possible to use it as a bilingual in any chosen variant. Later, the dictionary was supplemented with information about intralingual lexical relationships.

2.2 Computer model of ITB based on lexically related components
The information and terminological base of ML-technology is based on the results of the analysis of the language material. Linguistic material here should be understood as a certain set of texts of the analyzed subject area of a particular language. The size of the language material may vary depending on the analysis tools, the availability of original texts and the required number of terms.

The basis for the construction of ITB are EFD, according to which the priority of certain terms is determined. The use of such EFDs, obtained by analyzing linguistic material, qualitatively improves the process of analyzing foreign vocabulary in the study of the uniqueness of the text.

The ITB structure for ML-technology was developed through data-oriented structural analysis methodologies taking into account the frequency characteristics of its elements [18-19].

The terminological set corresponding to the basic information component of ML-technology was described in [41-43] as:

\[
\text{ML-component} = \{\text{language_term}_1, \text{language_term}_2, ..., \text{language_term}_N, \text{frequency_language}_1, \text{frequency_language}_2, ..., \text{frequency_language}_N\}
\]

The DSSSD methodology [44, 45] uses a similar notation, namely the multiple brace, as shown in figure 1.
Figure 1 - Description of the basic information component using DSSD notation.

Thus, the ITB will be represented by a set of objects, the structure of which is schematically shown in figure 2. Such combined objects are called lexically related components (LR-components) of the ITB [46-48]. The lexeme associated with all the lexemes of the LR-components of the ITB without exception is called the main lexeme. Lexemes with only one link are linked lexemes.

Lexeme types: 1 – main lexeme; 2, 3, 4, 5 – related lexemes; 1-2, 1-3, 1-4, 1-5 - lexical connections.

Quantitative characteristics: \( v_i \) – absolute frequency of the \( i \)-th lexeme; \( \mu_{ik} \) – absolute frequency of combination \( i \)-th and \( k \)-th lexemes.

Thus, the ML-component will no longer play the role of the basic information component. But this role will be played by the lexically related component. Let’s describe its structure. Let’s designate the main ML-component based on the main lexeme as MML-component, and ML-component based on the related lexeme as RML-component. Next, we will use the description of ML-component already known to us, and represent the terminological set as follows:

\[
\text{LR-component} = \{\text{MML-component (main lexeme), RML-component (related lexeme # 1), RML-component (related lexeme # 2), \ldots}\}.
\]

Figure 2 - LR-components in the ITB structure.

In DSSD notation (see [44, 45]) LR-component will look like it is shown in figure 3.
But at the same time, the structure of the MML-component based on the main lexeme (main) must be changed, otherwise information about lexical links is lost, and this contradicts our method. That is, in this case, the MML-component structure will look like this:

\[
\text{MML-component} = \{\text{language_term}_1, \text{language_term}_2, \ldots, \text{language_term}_N, \text{frequency_language}_1, \text{frequency_language}_2, \ldots, \text{frequency_language}_N, \text{combination}_1_{language_1}, \text{combination}_1_{language_2}, \ldots, \text{combination}_K_{language_N}, \text{frequency}_1_{language_1}, \text{frequency}_1_{language_2}, \ldots, \text{frequency}_K_{language_N}\}.
\]

Figure 3 uses DSSD notation to describe the MML-component. At first glance, the structure of the MML-component looks unwieldy. The question arises as to how resource-intensive it will be to form and operate such components. But on closer inspection, it becomes obvious that MML-components included ML-components, formed from text phrases. Thus, the initial volume of ITB has not changed, just as the resource intensity of its formation has not changed either. The resource intensity in the process of text analysis will increase slightly, since the analysis process is iterative, and there is only one such element at each iteration of the textual material analysis. The structure of the other ML-components (related) has not changed.

Lexical combinations and their frequency characteristics obtained by analyzing the neural network model of ITB [49] were included in the structure of the MML-component in order to preserve the multilingual approach. The fact is that for a phrase in a foreign language (at the moment of establishing an associative field), according to this technique, an analogue in another (foreign) language must be formed. This is nothing more than a linguistic analogue of this phrase. But in view of the different syntax of the languages included in the ITB, it is completely unacceptable to generate such language analogs by a simple combination of lexemes. Therefore, in order to preserve the principle of multilingualism, it is necessary to have exact linguistic analogues of lexical combinations as separate concepts of the subject area.

2.3 Analysis of text uniqueness based on lexically related components

The technique is based on the joint application of two algorithms [11]:

- the basic algorithm from [14], with the exception that the elements of the textual material are not lexemes, but LR-components [48];
- an algorithm for constructing intralingual associative fields [42].

The algorithm for constructing intralingual associative fields is simple and consists in consistently submitting elements of the LR-component to the analysis. The sequence of the algorithm steps is as follows:

- main lexeme → translation, text block hint (TBH) can be used in a foreign language (within ML-technology);
• related lexeme → translation, TBH can be used in a foreign language (within ML-technology);
• lexical combination of main and related lexemes → combination translation, TBH can be used in a foreign language (linguistic analogue of the lexical combination, but not lexemes separately);
• go to the next related lexeme;
• go to the next LR-component.

Using this technique, the basic algorithm will be significantly modified. First, and as already mentioned, the main element of the analyzed information will not be a term, but a lexically related component – LR-component, therefore, it becomes necessary to develop a mechanism for forming portions of text information, which would ensure the integrity of LR-components as basic. Secondly, based on the algorithm for constructing intralingual associative fields, we can say that the dependence of lexemes both in one language field and in different languages will be heterogeneous, since the emphasis is on the main lexemes.

Thus, the computer model proposed by the authors is the basis for modifying the algorithmic support for the process of analyzing the text uniqueness of based on lexically related components.

3. Conclusion
The solution to the problem of developing a methodology for analyzing the uniqueness of texts based on lexically related components is based on the following main conclusions:

• in computer modeling, electronic frequency dictionaries are the basis for information support for multilingual technology for analyzing the uniqueness of texts based on lexically related components;
• the development of multilingual electronic frequency dictionaries must be carried out in accordance with the trends in demand in the markets for text verification services for uniqueness and in accordance with the directions and problems of scientific, technical research, education and communication (materials of articles, monographs, dissertations) necessary for the development of key areas of scientific and educational activities in the field of preserving and increasing scientific and technical potential;
• an information-terminological basis as a set of lexically related components can be built on the basis of electronic frequency dictionaries with data both on intra-lingual lexical relationships and at the level of interaction between different languages;
• the method of multilingual text analysis based on lexically related components requires a modification of the basic algorithm and the development of algorithms for forming an information and terminological basis as a set of lexically related components.

Acknowledgements
The first version of the trilingual (English-German-Russian) dictionary was created in cooperation with colleagues from the Institute of German Studies at the Ruhr University in Bochum, Germany. The main work related to checking and correcting the German-language part of the dictionary was carried out by Professor I.V. Kovalev in collaboration with Professor H.-R. Fluck at the Ruhr University in Bochum (Germany). This collaboration was made possible with the financial support of the Frankfurt Airport Research Foundation (Flughafen Frankfurt-Main Stiftung) and the German Academic Exchange Service (DAAD).

References
[1] Apalkov V G 2013 Open Education 6(101) 71-5
[2] Artemova G O et al. 2015 Open Education 5(112) 4-10
[3] Atanov G A 2001 Educational Technology & Society 4(1) 111-24
[4] Atanov G A 2001 Educational Technology & Society 4(4) 48-55
[5] Galeev I H and Galeeva N H 2012 Educational Technology & Society 15(1) 594-608
[6] Galeev I H 2013 Educational Technology & Society 16(4) 387-402
[7] Galeev I H 2015 Educational Technology & Society 18(4) 767-79
[8] Avdeeva N V et al. 2018 Open Education 22(5) 74-83
[9] Andrianov I A and Grigorieva I A 2016 Effective search for plagiarism in the program code for the remote programming workshop system International Scientific and Practical Conference INFORINO-2016 485-8
[10] Absaliamova R A 2016 Open Education 3(19) 365-75
[11] Leskov V O 2008 Modern high technologies 12 29-30
[12] Leskov V O 2008 Modern high technologies 4 78-9
[13] Leskov V O 2009 Modern high technologies 2 63-4
[14] Leskov V O 2009 Modern high technologies 4 53-4
[15] Rastrigin L A 1981 Adaptation of complex systems (Riga: Zinatne)
[16] Skakunova V A 2013 Open Education 5(100) 24-7
[17] Smirnova V N 2019 Open Education 23(2) 4-13
[18] Kovalev I V et al. System aspects of multilingual adaptive-training technology organization and usage Proceedings of Modelling and Simulation MS’2004 - AMSE International Conference on Modelling and Simulation (Lyon-Villeurbanne France 5-7 July 2004) 14.11-13
[19] Kovalev I V et al. 2019 IOP Conf. Ser.: Mater. Sci. Eng. 537(4) 042085
[20] Sychov O A and Mamontov D P 2014 Open Education 2(103) 79-88
[21] Terentieva I A 2014 Actual problems of the humanities and natural sciences 11-2 185-7
[22] Trembach V M 2013 Open Education 4(99) 52-62
[23] Merchelevich G V and Zacharova O A 2019 Educational Technology & Society 2(22) 38-43
[24] Filatova N N and Achremchik O L 2000 Educational Technology & Society 3(2) 333-9
[25] Harina M V 2018 Open Education 22(5) 65-73
[26] Jablouchkin L B et al. 2001 Information Technologies 3 25-27
[27] Carroll J B et al. 2010 Frequency Dictionary of Contemporary American English: Word Sketches, Collocates, and Thematic Lists (Routledge)
[28] Eisenreich G und Sube R 1996 Langenscheidts Fachwörterbuch Mathematik: Englisch-Deutsch-Französisch-Russisch (Langenscheidt)
[29] Ferretti V 1996 Wörterbuch der Datentechnik: Englisch-Deutsch, Deutsch-Englisch; 9200 Einträge aller Gebiete der Informatik, mit Hauptbegriffen der angrenzenden Fachgebiete sowie des allgemeinen technischen Sprachgebrauchs (Berlin: Springer)
[30] Ferretti V 1992 Wörterbuch der Elektronik, Datentechnik und der Telekommunikation: Deutsch-Englisch (Berlin: Springer)
[31] Goodfellow R 1995 Computer Assisted Language Learning Journal 8(2) 205-26
[32] Haberfeller K and Daenzer W 2007 Systems engineering: Methodik und Praxis (Zürich: Verl.Industrielle Organisiation)
[33] Herdan G 1964 Quantitative Linguistics (London)
[34] Jones G 1986 System 17 35-47
[35] Kotik M 1986 Wörterbuch Luft- und Raumfahrttechnik: Russisch-Englisch-Deutsch (Düsseldorf)
[36] Leech G P and Rayson A W 2001 Word Frequencies in Written and Spoken English: based on the British National Corpus (London: Longman)
[37] Legenhausen L and Wolff D 2000 System 18(1) 1-13
[38] Levy M 2007 Computer Assisted Language Learning: Context and Conceptualization (Oxford: Clarendon Press)
[39] Kovalev I V et al. 2000 Modelling, Measurement and Control D 21(3-4) 11-26
[40] Pozdneev B M and Sytiagin M V 2015 Open Education 1(108) 4-11
[41] Karaseva M V et al. 2009 Control systems and information technology 1.3(35) 360-3
[42] Karaseva M V et al. 2009 Control systems and information technology 4(34) 78-82
[43] Kovalev et al. 2009 Software products and systems 4 91-4
[44] Kalyanov G N 1996 *CASE structural and systems analysis: automation and application* (Moscow: Publishing house "LORI")
[45] Kalyanov G N *et al.* 1996 *PC WEEK/RE* 34 46-50
[46] Karaseva M V 1994 *English-Russian frequency dictionary for system analysis* (Krasnoyarsk: SAA)
[47] Karaseva M V *et al.* 2009 *Software products and systems* 3 35-8
[48] Karaseva M V *et al.* 2008 *Control systems and information technology* 3.1(33) 157-60
[49] Usachev V and Kovalev I V 2003 Adaptive technology of the education quality control *Proceedings of the 9th International Scientific and Practical Conference of Students, Postgraduates and Young Scientists - Modern Techniques and Technologies - MTT-2003* (Tomsk 7-11 April 2003) 251-3