CEUR-WS.org/Vol-2523/paper29.pdf

The Software Environment for Multi-aspect Study of Lexical Characteristics of Text

Elena Sidorova and Irina Akhmadeeva

A.P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Acad. Lavrentjev avenue 6, 630090 Novosibirsk, Russia
{lsidorova,i.r.akhmadeeva}@iis.nsk.su

Abstract. The software environment for multi-aspect study of the lexical characteristics of the text is considered. The proposed environment provides tools and features allowing automatically building a dictionary based on a text corpus of interest. The created toolkit focused on lexical units acting as markers and indicators of higher level objects. The considered environment allows solving various text analysis tasks; because it integrates various tools for conducting language research and supports customization of vocabularies to a problem area. This toolkit includes interfaces for developing vocabularies and a system of features. To study the contexts of the use of terms, concordance construction tools are provided. Concordances allow the researcher to test his or her hypothesis about the functionality of a particular lexical unit. To describe more complex constructions to be extracted, a user can apply search patterns, supported by a user-friendly language. Using these patterns allows us to develop lexicographic resources containing not only the traditional vocabularies and stable inseparable lexical phrases, but also language constructs that have a more complex structure.

Keywords: domain vocabulary, terminology, concordance, search pattern

1 Introduction

A plain text, as it is a source of information, and one of the most important means for communication needs to be thoroughly studied. It is necessary for both evaluating "quality" of what has been written and automatic text processing along with supporting information retrieval services. Studying language phenomena and modeling text understanding processes taking place at the different language levels are in the focus of contemporary research in computational linguistics.

In order to work out these problems, it is usual to apply a variety of knowledge written in a formalized form. Among them are widely known thesauri such as WordNet and RusNet, explanatory combinatorial dictionaries, annotated corpora of texts (for example, The Russian National Corpus www.ruscorpora.ru), and other resources. Serving as an instrument for describing a subject vocabulary, thesaurus allows us to characterize terms and their connections from the point of view of peculiarities of use in this subject domain [1]. Another way of studying the linguistic phenomena is to use corpora of texts. A text corpus is the source and tool of multi-aspect lexicographic works [2]. The use of specialized methods, such as a frequency analysis of a vocabulary in the corpus,

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

construction of concordances on various grounds, can help in automating the work of experts on a formal structures research, initial filling of dictionaries, and construction of linguistic models on the basis of an annotated corpus of texts. Despite the widely demanded functionality, there are no known analogues of the specialized set of customizable components that integrate lexicographic research methods for Russian and provide semantic markup of terms, statistical analysis, and construction of concordances. As for other languages, similar functionality is presented on such platforms as GATE (https://gate.ac.uk) or CLARIN portal (www.clarin.eu). Components developed by various groups of researchers from different countries and for different languages are presented in these resources. As well as a method of integrating components into a chain of calculations is proposed.

Literature overview [3–5] shows that many researchers having a task to extract terminology from a large text collection usually choose to combine linguistic and statistical methods. For extraction of lists of candidate terms that satisfy the specified linguistic conditions, the method of search patterns describing classes of language expressions is used. Depending on the type of language information taken into account, the patterns used in various works are divided into grammatical, lexico-grammatical [3, 5] and lexico-syntactic patterns [7, 8]. Extraction of candidate terms is accompanied by calculation of statistics and weights for filtering and sorting a result list. The list of candidate terms includes not only special concepts established in this field, but also numerous general scientific, peripheral and author's terms that, as shown in [12], are characterized by a high degree of variation of the language form. In this situation, an expert assessment stage is needed, at which the ranked lists are presented to the expert for selecting true terms.

This paper concerns describing various supporting tools for studying lexical characteristics of a text based on corpora. Combining proposed tools allowed us to develop an environment for creating problem-oriented vocabularies and provide the end user with various possibilities to study language phenomena.

2 Requirements for the Text Research Support Environment

The development of linguistic models and the creation of resources of sufficient quality require scrupulous manual labor, supported by software tools. The software environment should provide the expert with various tools to create necessary knowledge bases and carry out case studies.

We formulate the requirements for the system for multi-aspect study of lexical characteristics of text as follows:

1. The system should be able to automatically fill vocabularies based on text corpora;
2. The system user should be able to customize and add various attributes for vocabulary terms;
3. The system should be able to carry out lexical analysis (segmenting a text, and extracting terms that are presented in the vocabulary);

4. The system should keep statistical and combinatorial properties of language phenomena found in texts;
5. The system should be able to build a concordance of terms and provide the user with corresponding visualization tools.

We developed a system including basic research tools that follows (Fig. 1): an interface for developing a dictionary and creating a group of features, tools for automatic generation of lexical content of a dictionary by the corpus of texts and calculating quantitative characteristics of found terms, concordance construction tools for studying contexts of lexical units.

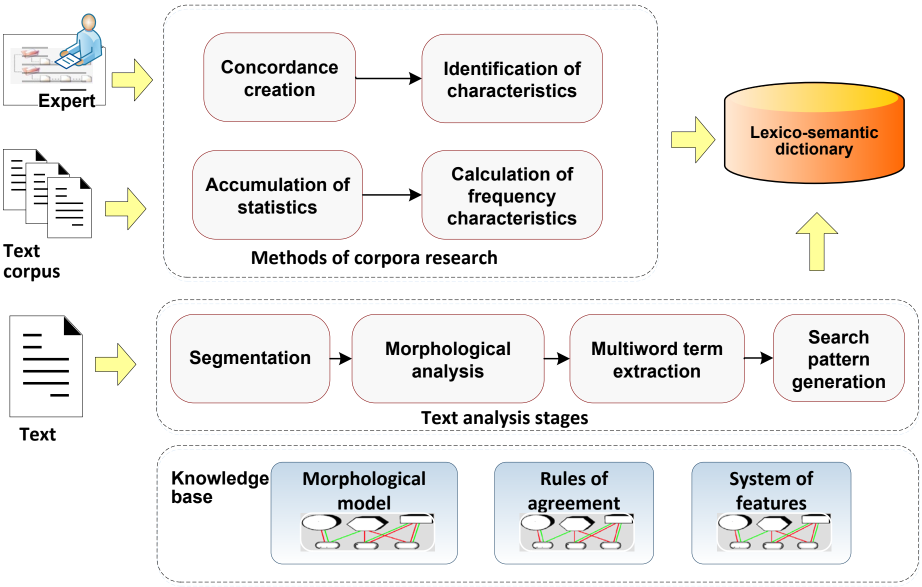

Expert
Concordance creation
Identification of characteristics
Text corpus
Accumulation of statistics
Calculation of frequency characteristics
Methods of corpora research
Lexico-semantic dictionary
Text
Segmentation
Morphological analysis
Multiword term extraction
Search pattern generation
Text analysis stages
Knowledge base
Morphological model
Rules of agreement
System of features

Fig. 1. The environment for study of lexical characteristics of text

3 Knowledge Representation Model

The considered lexicographic knowledge model includes three main components. The dictionary defines the lexical model of the sublanguage under consideration, which defined by the problem area. Grammar provides search and retrieval of lexical units from texts. The set of user-defined pragmatically-oriented features supports recording of observations, and is focused on further support of automated text processing methods.

A representative problem-oriented corpus of texts lies in the research basis. The main tools providing research support are following:

- search for examples of using vocabulary terms;
- build a variety of contexts (concordances);
- calculate frequencies, co-occurrences, distributions, etc.

3.1 Lexical Model

In our approach, the dictionary entry contains all information that necessary either for extracting terms from text or for supporting the subsequent stages of the text analysis.

A problem-oriented dictionary is a volume of vocabulary organized according to a semantic (thematic/genre/etc.) principle, considering a certain set of basic formal relationships. Formally, the dictionary is defined as a system of the form:

$$V=\{W, P, M, G, S, F_w, F_p\},$$

where W is a set of lexemes, where each lexeme is mapped to the entire set of its lexical forms; P is a set of multiword terms defined as a pair of a form <N-gram, structure type>. The N-gram specifies sequence of lexemes, and the structure type defines the head of the phrase and rules for matching N-gram elements.

M is a morphological model of language. It defines morphological classes and features.

G is a set of agreement rules which are used to extract multiword terms.

S is a problem-oriented set of features, terms could be marked with.

$F_w=W \to 2^{M \times S}$, $F_p=P \to 2^{G \times S}$ is a function that maps terms to sets of features.

The morphological representation the system provides is designed in such a way that it could be customized depending on the specific problem the user is working on. He or she can define his or her own set of features and classes, and ensure they are integrated in the basic morphological representation. A morphological class is defined by a part of speech, a set of lexical features (for example, animacy or gender for nouns) and a type of paradigm. It is a rather rare case when one would need to change class. For example, it is necessary when using additional specialized dictionaries of terms (dictionaries of names, geographical locations) or there is a need to include words of another language in the dictionary.

The description of morphological information includes the following concepts: morphological attribute, class, part of speech, and type of paradigm.

The morphological attribute is described by the name N_i and the set of its values X_i: $<N_i, X_i>$ (for example, <*Gender, {masculine, feminine, neuter}*>). Part of speech is also an attribute, but since it must always be present, it was decided to create a separate entity for this purpose. Attributes within each class are divided into derivational, inherent to all forms of the lexeme of this class, and inflectional, distinguishing forms of one lexeme.

The paradigm type determines its length and matches each element of the paradigm with a set of attribute values (for example, for a "simple" adjective it is a triple <case, number, gender>). Such elements are strictly ordered, which makes it possible to use a compact form of writing in a tree-like structure, the vertices of which are subsets of the attribute values $<A_i, X_i>$. A pair of functions $f: n \to X_{i1} * ... * X_{ik}$, $g: X_{i1} * ... * X_{ik} \to n$ provides a conversion of the inflectional paradigm to a set of attribute values, and vice versa. So each lexeme is assigned a paradigm from the paradigm table, and each paradigm is assigned a type of paradigm describing its structure.

The morphological class includes a part of speech, a set of derivational lexical features $x_{ij} \in X_i$ (for example, animation or gender in nouns) and a type of paradigm describing attributes of word forms.

Another important feature of the system is the support of multiword terms (phrases) formed according to the shallow syntactic analysis based upon a fixed set of rules. Most of the multiword terms include from two to four words and are formed using the rules of the following type:

- A+N (“*аналоговый датчик”* which means *“analog sensor”* in Russian) – agreement of a noun and an adjective;
- N+Ngent (“*автор учебника”* which means *“textbook author”* in Russian) – agreement of a noun and a noun in the genitive case;
- A+A+N (“*новая информационная технология” – “new information technology”*);
- N+Agent+Ngent (“*обработка естественного языка” – “natural language processing”*);
- A+N+Ngent (“*локальная степень вершины*” – “*local degree of a vertex*”),
- N+Ngent+Ngent (“*компонента связности графа*” – “*connected component of a graph*”) etc.

There are also terms with a more complex structure, for example, with dependent prepositional groups:

- N+PREP+N (“резервуар с жидкостью” – “reservoir with liquid”, “рассуждение по умолчанию” – “*default reasoning*”);
- N+PREP+N+N (“*поиск в пространстве состояний*” – “*search in the state space*”);
- N+PREP+A+N (“*автомат с переменной структурой” – “variable-structure automata”*) etc.

The system has its own component of multiword term extraction of the Russian language, which, according to a given set of words and their grammatical characteristics, checks agreement in accordance with one of the syntactic models and synthesizes a normal form of a multiword term. The multiword vocabulary term is uniquely identified by a triple <normal form, rule, < lexical structure>>. Such term has a syntactic head (a single-word term) and grammatical features that are formed on the basis of the grammatical features of the head.

3.2 Features of Terms

Depending on the problem being worked out, terms in the dictionary can be supplied with features of various types: statistical (for solving classification problems), genre (for text genre analysis), semantic (for semantic analysis), formal (for identifying markers of certain structures), etc.

Statistical features keep frequency information. When text is processed all terms occurred in it have their statistics updated. To perform text classification, we need a training corpus, i.e. corpus annotated with predefined set of interrelated topics. In the dictionary for each term we know how much times it occurred in the training corpus (this is called absolute frequency), and a number of texts in which the term occurred (text frequency). We also know a list of topics where term was found, absolute frequencies and text frequencies for each topic from the list. Some parameters (relative frequency, tf*idf, weight) are calculated dynamically.

The set of features user needs to markup dictionary terms with, are defined by him or her and depends on the task being addressed, so it is completely user-defined and problem oriented. To encode various information about the term (semantic, genre, stylistic, etc.), the following facilities are provided.

- Class. The term could be of one of the classes. A class hierarchy allows user to assign a term to a certain level of hierarchy: more general or specific, inheriting properties from upper classes.
- Attribute. Attributes are used to represent the lexical meaning of a term. Combining word's semantic attribute values, we can, to a certain extent, model the component semantic structure of a word. The main components of the semantic structure of the term can be considered as thesaurus descriptors.
- Alternative feature sets allow the term ambiguity to be expressed.

4 Working with Text Corpora

The developed environment consists of vocabulary components and processors that, on the one hand, allow automatic creation, fill and edit dictionaries, and, on the other hand, use those dictionaries in lexical text analysis. One of the most important features is user supporting tools such as term sorting, term filtering, text coverage visualization, concordance constructor, etc.

4.1 Corpus-based Vocabulary Learning

The terminology extraction process consists of steps that follows: a) text tokenization b) lexical and morphological analysis (lemmatization, extraction of lexical and grammatical features, normalization), c) extraction of phrases that are "look like" terms (phrase term-likeness is based on predefined grammatical models), d) update the statistics of found terms.

Following are the modules that are used for dictionary construction.

The morphological analysis is carried out on the basis of the Dialing module (www.aot.ru), which contains a dictionary of general Russian terms. This module supports search for words, along with their grammatical features and normal forms on the dictionary. It also provides an additional feature called predictor, which for any word that is not in the dictionary can make assumptions about part of speech, normal form and other features. Predictor can make up to three assumptions for a single term.

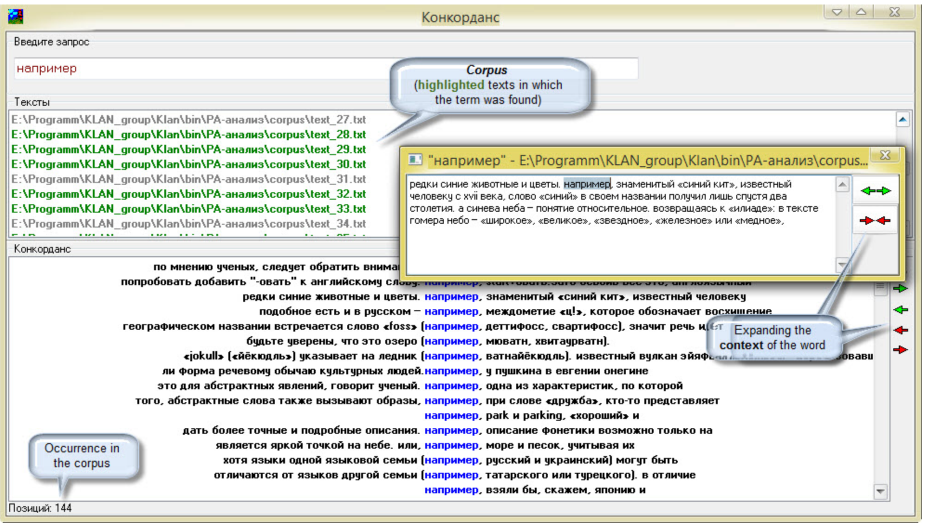
The multiword term extractor is applied to recognize phrases in accordance with a fixed set of grammatical rules. The main objective of the module is to identify the most important term-forming syntactic groups, most of which are nominal groups or are based on them.

Using aforementioned modules to process text corpus, we will end up with the resulting dictionary and statistics of frequencies of terms. If there were special features marked in the corpus, the corresponding terms are treated as having those features, and statistics are also kept with regard to the features.

The proposed environment hereby provides tools and features allowing automatic building a draft dictionary from scratch based on a text corpus of interest. On the basis of such a dictionary a further research could be carried out.

4.2 Concordance

A concordance is the traditional way of studying a corpus of texts. It contains a complete index of terms that share context with the selected one. The sizes of contexts may vary. Concordances allow the researcher to test his or her hypothesis about the functionality of a particular lexical unit. It could be said that a concordance connects dictionary terms with the text corpus, and serves as a linguistic markup at the morphological and shallow syntactic level.



Конкорданс
Corpus (highlighted texts in which the term was found)
Expanding the context of the word
Occurrence in the corpus

Введите запрос

например

Тексты

E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_27.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_28.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_29.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_30.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_31.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_32.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_33.txt
E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus\text_34.txt

"например" - E:\Programm\KLAN_group\Klan\bin\PA-анализ\corpus...

редки синие животные и цветы. например, знаменитый «синий кит», известный человеку с xvii века, слово «синий» в своем названии получил лишь спустя два столетия. а синева неба – понятие относительное. возвращаясь к «илиаде»: в тексте гомера небо – «широкое», «великое», «звездное», «железное» или «медное»,

Конкорданс

```
                                   по мнению ученых, следует обратить внима
                      попробовать добавить "-овать" к английскому сло
                                       редки синие животные и цветы. например, знаменитый «синий кит», известный человеку
                                          подобное есть и в русском – например, междометие «ц!», которое обозначает восхищение
             географическом названии встречается слово «foss» (например, деттифосс, свартифосс), значит речь и
                                        будьте уверены, что это озеро (например, мюватн, хвитаурватн).
                          «jokull» («йёкюдль») указывает на ледник (например, ватнайёкюдль). известный вулкан эйяф
                    ли форма речевому обычаю культурных людей.например, у пушкина в евгении онегине
                        это для абстрактных явлений, говорит ученый. например, одна из характеристик, по которой
                  того, абстрактные слова также вызывают образы, например, при слове «дружба», кто-то представляет
                                                                 например, park и parking, «хороший» и
                              дать более точные и подробные описания. например, описание фонетики возможно только на
                                  является яркой точкой на небе. или, например, море и песок, учитывая их
                                хотя языки одной языковой семьи (например, русский и украинский) могут быть
                               отличаются от языков другой семьи (например, татарского или турецкого). в отличие
                                                                 например, взяли бы, скажем, японию и
```

Позиций: 144

Fig. 2. The concordance construction tool

The implemented in the environment concordance construction tool works with text files. The user can customize the size of a text fragment being viewed in a context of a term entry (Fig. 2.). An example of a concordance given in Fig. 2 for a word "*например*" (which means "*for example*" in Russian) includes 144 occurrences from the text corpus, and shows how context could be expanded word-wise, or how one or more paragraphs could be summoned to view by providing the selected term entry. In the example the research purpose was to test the hypothesis about the use of this term in the *argument from expert opinion*.

In general, this kind of research allows user to identify more complex language constructs that ensure the precision and recall of the information extraction process, and to identify additional features based on them. To describe constructions to be extracted, user can use search patterns which based on regular expressions, supported by a user-friendly language [7, 8].

4.3 Search Patterns

In our studies, we have been using different types of patterns and tools that support automatic text processing. In each case the toolkit was chosen based on a problem area and methods used to solve the target problem.

For example, in the project targeting the problem of filtering out prohibited content [9], in addition to being marked by thematic, genre and lexico-semantic features got from the vocabulary texts was processed with special patterns each of which described constructions specific to a particular Internet genre [10]. Those patterns have significantly improved the accuracy of the genre classification.

Taking a closer look, a pattern allowing detecting a block containing personal information on a website can be represented as follows:

_profile:[“личный кабинет”][“профиль”][“аккаунт”][“о себе"][“личный профиль”]

//_profile: [“personal account”] [“profile”] [“account”] [“about me”] [“personal profile”]

Profile Description / Contacts: [<_profile, all_h>]

In this case, the *_profile* pattern is defined by a set of alternative terms. If any of these terms appears as a part of a header at any level (as indicated in the second pattern) we can classify a text block as a block containing user profile information. Patterns defined like one from the example belongs to logical combinatorial lexical patterns.

In another project our goal was to extract information from technical documentation texts. We built a glossary of terms with semantic subject-oriented markup, and applied search patterns to extract parametric information, which is often represented by numerical and symbolic notations and abbreviations. The patterns used are defined as follows.

class: ‘Object ACS’, template: ‘ACS TP’, type: ‘base’

[АСУ] = АСУ{ТП}; автоматизированн{...} систем{...} управления

[ACS] = ACS{TP}; automatic control {...} system

Patterns of this type are called lexical-semantic patterns [11].

Finally, there is yet one project where we target a philosophical problem of argument analysis. We build a dictionary of markers of argumentative structures on the basis of an annotated text corpus. Applying patterns allows us to represent area specific constructions, which could be consisting of more than one part, separated with gaps.

DSC = [begin: DS, w / <speech> <Verb, past | present>, Expert <N, им>, end: ES]

quote_l = [“|«]

quote_r = [”|»]

DS = [begin: quote_l, end: quote_r] // direct speech

Thus, in the experiment on the extraction of arguments from expert opinion, the search accuracy using patterns was 86.5%. Based on the above we can conclude that using our search patterns allows us to develop lexicographic resources containing not only the traditional vocabularies and stable inseparable lexical phrases, but also language constructs that have a more complex structure.

5 Conclusion

This paper is devoted to describe approaches and methods for development of lexicographic resources, conducting studies on text corpora in order to ensure the completeness and reliability of models being developed. The created toolkit is focused on lexical units acting as markers and indicators of higher level objects (semantic, pragmatic, structural-genre, logical-argumentative, etc.).

The considered software environment integrates basic tools required to conduct research on lexical characteristics of the text, which ensures a full cycle of the expert's work. The environment has wide possibilities for tuning of parameters, ranging from grammatical categories, lexico-semantic characteristics, classification parameters, and ending with specific search patterns that ensure the search for contexts and the construction of concordances. Practical use of this software in various research projects showed usability, the relevance of functionality and adaptability for different tasks.

Consequently, distinctive features of the system are:

- possibility of multipurpose use in solving various text analysis tasks, such as text classification, information extraction, lexicographic research of a text corpus, genre analysis, etc.;
- integration of various tools within the same environment for conducting language researches and providing customization of vocabularies to a problem area: concordance, statistical study based on a corpus of texts, support for semantic markup of lexical units, a rich set of search tools and filtering.

The environment supports a rich lexical model that integrates various models of representation of lexical units and language constructs.

1. Expandable and customizable morphological model (in contrast to the well-known morphological analyzers aot, pymorphy, mystem, etc.);
2. Grammar models for Russian phrases extraction and the possibility of selectively use them;
3. Search patterns integrate semantic, grammatical, lexical and symbolic representations based on logical operations.

Further improvements of the system may lie in developing of corpus-based research tools, such as constructing concordances for the joint occurrence of terms, using conditions for the presence / absence of feature sets in search queries, etc. It is also planned to enhance the reusability of results of the research by storing the data in standard formats based on XML (TEI, OWL).

Acknowledgment

The work was carried out with the financial support of the Russian Foundation for Basic Research (grants 19-07-00762 and 17-07-01600 and 18-00-01376 (18-00-00889)).

References

1. Loukachevitch, N.V.: Thesauri in information retrieval tasks. MSU Publ., Moscow (2011).
2. Sinclair, J. Corpus, Concordance, Collocation. Oxford University Press, Oxford (1991).
3. Zakharov, V.P. and Khokhlova M.V.: Automatic extracting of terminological phrases. Structural and Applied linguistics **10**, 182–200 (2014)
4. Bolshakova, E., Loukachevitch, N., and Nokel, M.: Topic models can improve domain term extraction. In: International conference on Information Retrieval ECIR-2013, pp. 684–687. Springer Verlag, (2013)
5. Mitrofanova, O.A. and Zaharov V.P.: Automatic extracting terminological phrases. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog-2009", pp. 321–328. Moscow (2009).
6. Sokirko, A.V.: Morphological modules on the site www.aot.ru. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog-2004", pp. 559–564. Nauka Publ, Moscow (2004).
7. Bol'shakova, E.I., Baeva, N.V., Bordachenkova, E.A., Vasil'eva, N.E., and Morozov S.S.: Lexicosyntactic patterns for automatic text processing. In: Proc. Int. Conf. Dialogue 2007, pp. 70–75. Moscow (2007).
8. Rabchevsky, E.A., Bulatova, G.I., and Sharafutdinov, I.M.: Application of lexical-syntactic patterns to the automation of ontology building process. In: Proc. 10th All-Rus. Conf. RCDL'2008 Electronic Libraries: Perspective Methods, Technologies, Electronic Collections, pp. 103–106. Dubna (2008).
9. Sidorova, E.A., Kononenko, I.S., and Zagorulko, Yu.A.: An approach to filtering prohibited content on the web. In: CEUR Workshop Proceedings, 2022. pp. 64–71. CEURWS.org (2017).
10. Sidorova, E.A. and Kononenko, I.S.: Genre aspects of websites classification. Software Engineering **8**, 32–40 (2015).
11. Sidorova, E.A. and Timofeev, P.S.: A lexico-semantic templates as a tool for declarative description language constructs linguistic text analysis. System Informatics **13**, 35–48 (2018) DOI: 10.31144/si.2307-6410.
12. Bol'shakova, E.I. and Ivanov, K.M.: Term extraction for constructing subject index of educational scientific text. In: Sixteenth Russian Conference on Artificial Intelligence RCAI-2018. T1, pp. 253–261. Moscow (2018).