ISSN 0146-4116, Automatic Control and Computer Sciences, 2019, Vol. 53, No. 7, pp. 705–718. © Allerton Press, Inc., 2019.
Russian Text © The Author(s), 2018, published in Modelirovanie i Analiz Informatsionnykh Sistem, 2018, No. 4, pp. 435–458.

Russian-Language Thesauri: Automatic Construction and Application for Natural Language Processing Tasks

N. S. Lagutina[a], *, K. V. Lagutina[a], **, A. S. Adrianov[a], *, and I. V. Paramonov[a], ******

[a] *Demidov Yaroslavl State University, Yaroslavl, 150003 Russia*

**e-mail: lagutinans@rambler.ru*

***e-mail: lagutinakv@mail.ru*

****e-mail: alex.a4.25@yandex.ru*

*****e-mail: Ilya.Paramonov@fruct.org*

Received August 1, 2018

Abstract—The paper overviews the existing digital Russian-language thesauri and the methods of their automatic construction and application. The authors have analyzed the main characteristics of thesauri published in open access for scientific research, evaluated trends of their development, and their effectiveness in solving natural language processing tasks. Statistical and linguistic methods of thesaurus construction that allow automation of their development and reduce the labor costs of expert linguists have been studied. In particular, algorithms for extracting keywords and semantic thesaurus relations of all types have been considered and the quality of the thesauri generated with the use of these tools was assessed. To illustrate features of various methods of constructing thesaurus relations, the authors developed a combined method that fully automatically generates a specialized thesaurus based on a text corpus of a selected domain and several existing linguistic resources. The proposed method was used to conduct experiments on two Russian-language text corpora that represent two different domains: articles on migration and tweets. The resulting thesauri were analyzed by means of an integrated assessment that had been developed by the authors in a previous study and allows one to determine various aspects of the analyzed thesaurus and appraise the quality of the methods of its generation. The analysis revealed the main advantages and disadvantages of various approaches to thesaurus construction and extraction of semantic relations of different types, and also made it possible to identify potential focus areas for future research.

Keywords: thesaurus, semantic relations, automatic thesaurus construction, automatic relation extraction, keyword extraction

DOI: 10.3103/S0146411619070149

INTRODUCTION

A thesaurus is a vocabulary of natural language terms that includes a system of relations between these terms [1]. A thesaurus can be used both as a resource for information retrieval and as a terminology source or reference material. Relations between words serve as a material for constructing lexical–semantic networks for extracting knowledge and determining the words' semantic proximity. The formalized nature of thesauri makes it possible to easily automate their application. Many researchers emphasize the importance of constructing digital thesauri and the prospects of using them in systems of automatic text processing [2, 3].

While solving natural language processing problems, the authors discovered the fact that a thesaurus can act as a convenient model of a domain. The authors have successfully used an automatically generated thesaurus to construct an alternative navigation system for the e-tourism resource Open Karelia [4] and also for sentiment classification of newspaper articles [5]. It should be noted that these studies were conducted on English texts and utilized corresponding methods, algorithms, and linguistic resources.

The attempt to apply the developed approaches directly to analyzing Russian-language texts did not prove equally successful. Even the most superficial analysis demonstrated a low quality of the generated thesauri, in particular, a small number of extracted relations between words. Additionally, using an external digital thesaurus, RuThes, in the algorithms as an alternative to the English-language thesaurus WordNet did not affect the quality of the solution in any significant way. This prompted the authors to analyze

the methods of Russian-language thesauri construction and consider the existing digital Russian-language thesauri more closely.

In 2015, a brief analysis of Russian-language thesauri was performed in [6, 7]. The authors of the overviews note the insufficient volume of these resources, in particular in comparison with WordNet; the difficulty of integration with existing algorithms and natural language processing systems; and the complexity or impossibility of their use, particularly for commercial purposes. However, digital resources can change quite quickly, so it would be interesting to identify the dynamics and trends in the development of Russian-language thesauri.

The existing thesauri are not able to fully cover the whole range of natural language processing issues, especially ones specific to individual domains, so the task of constructing new thesauri—both as independent resources and as auxiliary tools—is of immediate interest. Unfortunately, the authors of this study were not able to find overviews of automated construction methods for Russian-language lexical resources. Therefore, describing the methods used for automatizing thesaurus construction and determining development directions for these methods was chosen as one of the goals of this article.

The presented study consists of three parts. The first part analyzes the characteristics of digital Russian-language thesauri, determines the dynamics of these resources' development, and assesses the possibilities of applying them to natural language processing tasks. In the second part, the authors consider the currently used methods for constructing lexical resources of extracting terms and relations between them to determine the main trends and possible development directions in this area, primarily in what concerns thesauri construction. The third part describes the authors' own experience of constructing Russian-language domain-specific thesauri. The conclusions section sums up the results of the study.

1. EXISTING DIGITAL RUSSIAN-LANGUAGE THESAURI: AN OVERVIEW

For English-language thesauri, the free linguistic resource WordNet (https://wordnet.princeton.edu) can arguably be considered a model one. The project has been developed since the early 1980s and is currently actively used in research. For the Russian language, several projects of this kind exist.

1.1. Properties and Characteristics of Thesauri

The largest thesaurus by volume is RuThes, a project developed by the Laboratory of Information Resources Analysis headed by N.V. Loukachevitch [8]. It is based on WordNet construction principles, but uses a different entity description model. Each unit of the thesaurus is a concept with an attached set of terms with values corresponding to the concept. The terms can be words or phrases and their number can be quite large, 20 or more. Words and phrases that are related to the same concept are termed as ontological synonyms.

Concepts are connected by a system of relations. RuThes uses four types of relations: two types of hierarchical relations—"class–subclass" and "part–whole"—and two types of associative relations—symmetrical association, which connects concepts that are very close in meaning, but are not joined into one concept, and unsymmetrical association, which connects two concepts that cannot exist without each other, but are not connected by any other relations.

At the moment the RuThes thesaurus contains 55 thousand concepts, 158 thousand words and expressions, and 210 thousand relations between these concepts. As is evident from Table 1, this amount of information is comparable to the volume of WordNet in all respects except the number of entities. However, the RuThes concept model and the WordNet synset model differ in meaning, so this quantitative comparison is not quite appropriate in this case.

An XML version of the thesaurus is available on request for noncommercial use. In addition, RuThes-lite, a free version of RuThes that contains 115 thousand words and expressions, can be modified and copied likewise for noncommercial purposes [9].

RuThes is developed and updated mainly by the experts from the Laboratory of Information Resources Analysis. This approach guarantees thesaurus quality, but necessitates considerable time for updating and modifying the data. The authors pay attention to the feedback about their project and take substantial comments into account. In 2017, RuThes was automatically converted to WordNet format in order to integrate with the existing global ontologies and thesauri in other languages. As a result, the RuWordNet thesaurus was created, containing 111.5 thousand Russian-language words and expressions [10].

Yet another active project of constructing a Russian-language thesaurus is Yet Another RussNet (YARN) [11]. The developers use a synset model that fully corresponds to WordNet synsets—groups of synonyms and quasi-synonyms, united by a shared lexical meaning. Synsets are linked together by hierar-

Table 1. Characteristics of Russian-language thesauri in comparison with the English-language thesaurus Wordnet

Characteristic	RuThes	YARN	Wordnet.ru	RussNet	Wordnet
Base unit	Concept	Synset	Synset	Synset	Synset
Quantity of units	55 k	70 k	20 k	5.5 k	117.7 k
Quantity of words, phrases	158 k	143.5 k	100 k	15 k	155 k
Relation types	Hierarch., associat.	Hierarch., antonymy	Hierarch.	Hierarch., antonymy, associat.	Hierarch.
Quantity of relation pairs	210 k	134 k	–	–	207 k
Usage format	XML	XML	Text files	–	API
Latest update	Continuously updated	Continuously updated	28.08.2008	14.06.2005	Continuously updated
Head organisation	Laboratory of Information Resources Analysis	Ural Federal University, Saint Petersburg State University	Unaffiliated research	Saint Petersburg State University	Princeton University

chical relations and relations of antonymy (established between terms with opposite meanings). Additionally, YARN also includes relations between words and interlingual links between YARN and WordNet synsets. Wiktionary data was used as a starter content source.

A distinctive feature of this project is the crowdsourcing approach to appending the thesaurus: any user registered on the YARN website can participate in adding and editing data. The authors of the project declare quality control by selecting a body of editors from the numbers of the website's users. The editors can approve or discard changes, as well as prohibit further editing of individual elements of the thesaurus. Thus, the quality of the project directly depends on the selected editors' qualification levels.

YARN currently contains 143508 words, 69799 synsets, 104906 unedited synonym pairs, and 29764 unedited hierarchical hypo-/hypernymic relations. The numbers are smaller than those for RuThes and WordNet, but a comparison of YARN's volume in 2015 [7] and 2018 demonstrates an evident growth pattern. That seems to be due to automatizing the process of extracting data from Wiktionary and similar resources and to the crowdsourcing approach.

An undeniable advantage of the project is the fact that YARN is distributed under a license that allows the use of the materials in research and commercial applications, as well as copying and modifying the data, provided that a reference to the source is included. The content of the thesaurus is provided in XML format.

Another open access thesaurus is Wordnet.ru or Russian Wordnet [12], an attempt of a translation of WordNet. The resource's construction system is fully automatic and does not presuppose expert evaluation. The authors used a heuristic algorithm of verifying the correspondence of the English synsets to the Russian synsets that had been obtained as a result of a direct translation. They managed to generate about 25 thousand synsets and 100 thousand words and phrases this way.

The authors have developed a data visualization program for Russian Wordnet, with attached text data that can be downloaded as a separate archive of text files. Unfortunately, the archive files were last modified on August 28, 2008. The small volume and the lack of development of this project provide evidence, firstly, of the complexity of building a fully automatic thesaurus construction tool and, secondly, of the infeasibility of constructing a full-fledged lexical resource for a language relying exclusively on translation, without taking national characteristics into account.

Another digital thesaurus that can be mentioned here is RussNet [13], developed at the Saint Petersburg State University's department of mathematical linguistics. The model of the resource is fully consistent with WordNet. The authors have constructed the thesaurus to be as compatible as possible with the European multilingual thesaurus EuroWordNet, which utilizes a system of interlingual references (Inter-Lingual-Index, ILI), which links terms of one language to words of another language that are similar, but not necessarily identical, in meaning. This system enables the use of the thesaurus for multilingual search. It should also be noted that this resource contains the largest number of relations of different types, thanks to the contributions of highly qualified expert linguists. Unfortunately, the closed nature of the project and the fact that it has not been updated since 2005 negates all its advantages. However, a project of integrating RussNet and YARN is currently underway [14]. That integration will not only increase the volume of the

combined thesaurus, but also improve its quality due to the fact that RussNet was created by a group of highly qualified experts.

As for open domain-specific thesauri, the authors only managed to find one project: the Russian version of the multilingual thesaurus of agricultural terminology AGROVOC [15]. The AGROVOC translation into Russian was performed by specialists from the Central Scientific Agricultural Library in 2011.

To sum up, at the moment we can name two Russian-language linguistic resources equivalent to the English WordNet—RuThes and YARN. The three resources are comparable in terms of data volume, both in the number of words and phrases and in the number of relations between them. The authors consider the advantages of RuThes to be the quality assurance provided by its team of experts and the reasonable approach to development and transformation. YARN's advantage is its compatibility with international lexical resources. In fairness, it should be noted that RuThes developers have created a similar compatible version, RuWordNet; however, it is still smaller than the source in volume. Another positive aspect of YARN is the possibility of its commercial use.

1.2. Applying Existing Thesauri to Specific Tasks

An analysis of digital Russian-language thesauri would be incomplete without reviewing studies that utilize these resources to perform specific tasks.

The existing thesauri are being used as a terminology source. The authors of [16] use RuThes as the basis for constructing a socio-political thesaurus of a national language (Tatar). In this case, RuThes acts primarily as an ordinary philological thesaurus-type dictionary. The authors emphasize the possibility of using this kind of lexical resources in various applications of automatic processing of news documents, legal acts, or social network posts. Therefore, the resulting thesaurus has been transformed and published in the Linguistic Linked Open Data (LLOD) cloud [17].

It should be noted that in order to represent the specific features of the Tatar lexical–semantic system, the authors actively use a large number of external sources. Such sources are, firstly, the existing bilingual Tatar–Russian dictionaries, including specialized socio-political ones, and, secondly, a large number of media texts and texts of official documents. The domain-related text documents are necessary for the purposes of finding Tatar terms suitable for replacing obsolete words or adding missing concepts.

RuThes is used as a terminology source in [18], which considers the problem of automatic text categorization. Each thesaurus concept is matched to a text category. The studied texts are searched for words that are also present in RuThes. The received data is used as the basis for thematic representation models constructed for each text and the identification of the text's category. The experiments produced high quality results. However, in order to deal with a category specific to a narrow domain, new terms had to be added to the thesaurus.

The structure of a thesaurus can be used as a template for other linguistic resources. The structure of RuThes has been used in this way for the construction of a security thesaurus [19]. This thesaurus is now used in a specialized data analytics system, where its uses include automatic text document classification. The data for the thesaurus is extracted from other sources: reference literature, specialized text collections, and mass media news articles on the subject of security.

Relations between words in thesauri are the key information used in methods of calculating terms' semantic proximity. Paper [20] makes an attempt to automatize the assessment of students' answers to open-ended test questions. An analysis of Russian-language answers was performed using RuThes and Wikipedia to calculate the semantic proximity of the words in the student's answer and the words in the reference answer. The positions of the words in the text were not considered. After analyzing the results, the authors found that the quality of the system's performance greatly decreased in cases where synonyms were involved, for example, when the reference answer contained duplicate possible answer versions. However, the authors only use class–subclass hierarchical relations and do not take synonymic relations between words into account.

All kinds of RuThes relations are used for calculating similarity of terms in [21]. Similarity of terms is one of the similarity factors in the developed method of constructing groups of semantically similar words and expressions that describe thematic nodes of a news cluster. The method was applied to the problem of automatic news abstracting.

A thesaurus could serve as a template for evaluating the quality of natural language processing methods. The author of [22] proposes a method of automatic grouping of semantically similar words. The results of the method's application were evaluated using RuThes and YARN materials. The quality of the results was not very high, especially when comparing with the RuThes thesaurus. That is due to the fact

that the author's method is based on using synonym graphs, and YARN synsets are much closer to the definition of a synonym than RuThes units (concepts).

The overview demonstrates that the thesauri are being used in a wide range of research (RuThes more frequently, YARN less frequently), but almost always in conjunction with other linguistic resources. The uses of these thesauri for domain analysis are very limited. In particular, the authors of [20] note the lack of domain-specific terms in the digital resources they use.

The multipurpose lexical resources RuThes and YARN contain useful information, are convenient to use, and also are developing dynamically. However, a huge number of automatic text processing tasks require domain-specific thesauri. Due to the complexity and labor intensity of constructing such resources, their number is quite small. However, the current level of the development of the methods of automated thesaurus construction makes this task feasible; in light of this, the authors would like to direct attention to the topicality of constructing open Russian-language thesauri for various domains.

2. FEATURES OF AUTOMATIC RUSSIAN-LANGUAGE THESAURI CONSTRUCTION

The task of constructing a thesaurus can be divided into two large subtasks: keyword extraction and relation extraction.

2.1. Keyword Extraction

Keyword extraction is the part of the thesaurus construction process (and of many other natural language processing tasks) that has been studied the most, for the Russian language as well. Statistical methods are of primary importance here. For the English language, keyword extraction methods, including text preprocessing, have been well researched and verified [24]. Similar methods are used for the construction of Russian-language thesauri; however, much less studies are dedicated to thoroughly assessing their quality.

In the formation of RuThes, an automated method of obtaining terminology was used [25]. For each of a selected number of domains (mathematics, physics, chemistry, biology, geology), collections of documents (consisting of 3000 to 8000 documents, 50 to 90 MB each) were formed. The collections' sources were documents available online: school lesson materials, abstracts, university lectures, and materials from specialized sites. In order to identify terms, two algorithms of extracting termlike words and expressions were used [26].

The first algorithm identified nouns, adjectives, pairs, and triples of nouns and adjectives in agreement, as well as predefined constructions (noun + noun in the genitive, etc.). The second algorithm identified frequently repeated words and multiword expressions. The resulting words and expressions were ordered by descending frequency order and descending number of documents that contain them. The list of keywords was verified and expanded by comparing it with the terms of the socio-political thesaurus, developed by the authors earlier.

It should be noted that the socio-political thesaurus and later RuThes have been developed by the research team since the late 1990s. The projects' main method of quality assessment is expert evaluation, which ensures high quality, but once again confirms the complexity and labor intensity of the work. The problems of natural language processing require developing methods for quicker construction and quality assessment of narrow domain-specific thesauri.

Most of the studies that describe automatic construction of own thesauri or similar structures (such as lexico-semantic networks or domain ontologies) extract terms using the words' frequency in the utilized text corpus and a measure of the terms' semantic proximity [27]. A word's frequency is determined by the number of its occurrences in the corpus. The semantic proximity measure is intended for quantitative evaluation of the terms' semantic similarity. This characteristic can be calculated in different ways and is often based on constructing a context multidimensional vector for each word and then analyzing the resulting vector space.

The author of [28] considers the problem of constructing a thesaurus, or rather a lexico-semantic field around a given term. The source material selected for the study was the term "engine" ("двигатель") and the ruTenTen 2011 corpus, which contains 14.55 billion words. The chosen statistical characteristics of the words were the degree of semantic proximity of the given word to the key word and the frequency of the given word in the corpus. The words were ordered by degree of semantic proximity. Expressions containing two or more words were also extracted, on the basis of a set of templates and of frequency calculations. However, the formation of the thesaurus was finalized by an expert.

Frequency is used as the sole basis for term extraction in [29]. That article investigates the application of a method of automatic construction of domain-specific thesauri for the construction of domain-specific ontologies to be used in education. The thesaurus was constructed based on a corpus of texts related to the domain. Unfortunately, it is not clear from the paper how the study identified the relations between terms. All stages of the formation of the thesaurus were evaluated by an expert.

Paper [30] proposes a method of calculating weighting coefficients in order to determine the degree of a given term's significance for the given domain. The authors consider a corpus of scientific articles, where they identify three numerical variables. The first one is frequency rank, which allows to equalize the significance values of the most common terms of any text and, at the same time, distribute the significance of terms within a single text. The second one is compliance with the text's topic. The topic groups of scientific texts were formed by selecting terms from the title and section headings in cases when the terms occur in the text itself with a high frequency; in these cases, the contribution to the value of the term's weighting coefficient is 1, in other cases it is 0. The third coefficient was calculated according to whether or not the term belongs to the conceptual blocks of the scientific article, as defined by the authors. Indicators and markers were identified for each of the four main blocks: problem, method, solution, results. If the term is used in a sentence that contains a formal indicator of one of the blocks, the term's weight is adjusted by the appropriate amount. Additionally, if the term is found in more than one block, its weight is adjusted by the sum of the corresponding values. The final assessment was calculated as an integral weighting coefficient of the term based on the values of the three variables. Thus, a system of metrics closely related to the structure of scientific papers was proposed for the analysis of text documents of that genre.

The described method was used in the development of a method of automatic topic identification for scientific texts [31]. That study demonstrates that for different domains both the methods in general and their individual details may vary significantly or change. Although the analysis of a document's structure is more often used in algorithms for extracting knowledge from texts, this approach is still justified for the construction of domain-specific thesauri in cases when the utilized text corpus or a part of it has a predefined format.

In [32, 33], the development of the Russian sentiment lexicon RuSentiLex is described. That linguistic resource was constructed mostly by hand, but some of its terms were extracted from texts posted on the social network Twitter using an algorithm that combines binary classifiers Logistic Regression, LogitBoost, and Random Forest. The accuracy of the terms extraction was assessed by means of comparison with an English review corpus, the terms of which had already been sentiment-tagged. It reached 78.6%. It should be noted that the quality of the algorithm's performance on Russian texts was assessed by experts manually, with no automation involved. The authors also proved the lexicon's practical applicability for sentiment analysis by organizing the SentiRuEval-2016 competition, where the participants used RuSentiLex to develop classification algorithms, the classification quality of which was in range of 55–68% for the F-measure.

We would like separately consider assessing the quality of keyword extraction. There are standard numerical characteristics used for this purpose: accuracy, completeness, and F-measure [24]. However, articles that deal with Russian-language texts rarely calculate these characteristics. Generally, the result is verified by an expert, most often without mentioning any numerical parameters except quantitative ones. Most likely, this is due to the lack of tagged domain-specific Russian-language text corpora, as noted in [34]. Authors of specific studies do not publish in open access the text corpora they developed and used. Of course, in many cases there are objective reasons for that, but the publication of even a small number of tagged Russian-language text corpora would significantly contribute to the development of data processing automation in this area.

In [35], the quality of the extraction of phrases as domain terms was evaluated with the use of the earlier developed Linguistic Ontology on Natural Sciences and Technology (SCI-Ontology). Average accuracy of selected terms was chosen as the measure. Its values were found to range from 59 to 75%, depending on features of the selected phrases. Unfortunately, this method can only be applied to the few domains for which available ontologies exist. Another problem with this method is forming a text corpus for the research.

It should be noted that the set of methods for keyword extraction used for analyzing Russian-language texts is rather small. According to the researchers' conclusions, the commonly used approaches and standard tools that can be applied irrespective or almost irrespective of the language provide a level of quality that is satisfactory for the considered problems. However, the authors postulate that the small number and the subpar quality of the existing digital Russian-language thesauri, especially in comparison with

English-language resources, is directly related to the lack of research into the entire range of the existing keyword extraction methods.

2.2. Relation Extraction

There are two types of semantic relation extraction algorithms: statistical and linguistic. The former analyze the frequency of terms' occurrences in texts to calculate their statistical characteristics and then apply mathematical methods to determine how semantically close a pair of terms is. In most cases, statistical methods identify associative relations. Linguistic methods are applied to existing available resources, such as dictionaries, thesauri, ontologies, etc., in order to extract predefined relations or apply linguistic rules or templates. Linguistic methods can identify all kinds of semantic relations. An overview of these methods can be found in the authors' earlier paper [36].

Researchers actively apply methods of automatic semantic relation extraction to English-language texts. The analysis of publications on the development of Russian-language linguistic resources immediately demonstrates a low degree of automation in methods of solving that problem.

RuThes developers devote a lot of attention to describing different types of relations that may be present in the thesaurus [37, 38]. However, during thesaurus constructing, RuThes relations were mostly carried over from the socio-political thesaurus and then supplemented by experts.

Experts are also in charge of relation extraction in a number of other studies [28, 29, 39]. Meanwhile, [28] clearly highlights the necessity of automatizing the construction of narrow domain-specific thesauri and the fact that relations between words are an important part of such thesauri.

Many authors rely on the relations between terms that are already described in available linguistic resources, mainly Wikipedia and RuThes. In [31], solving the problem of automatic topic identification of text documents involves constructing a specialized ontology, which uses synonymic and hierarchical relations from Wikipedia. Expert analysis of the results demonstrated an improvement in the quality of the system's performance when using the ontology, but the authors note the complexity of constructing such lexical resources. The use of RuThes was discussed in Section 2.

A method of extracting relations between words by using statistical methods based on cooccurrence frequency is applied in [40]. The researchers did not explicitly construct a thesaurus, but demonstrated that the problem of Russian-language text classification is solved better if semantic relations between domain terms are used.

The open distributional thesaurus of the Russian language [41] was built automatically using the Skip-Gram method implemented in the word2vec tool (https://code.google.com/archive/p/word2vec/), which identifies semantic relations between terms using statistical algorithms. The authors also automated thesaurus quality assessment by comparing the relations extracted for it with the existing resources that have tagged relations between words: the BLESS corpus, the corpus of cognitive associations, and others. The average accuracy of relation extraction reached as high as 97%.

In [42], a fully statistical method for automatic construction of associative relations is proposed for thesauri of several languages, including Russian. The method is based on calculating the words' cooccurrence, singular value decomposition of the term-document matrix, and several semantic proximity measures. For the English-language thesaurus, expert quality assessment demonstrated very high values of the metrics: accuracy and completeness reached 86–97%. It should be noted that quality assessment of the Russian-language thesaurus is not described in the paper in sufficient detail; specifically, the values of metrics that would describe the quality of the thesaurus as a whole are not given.

Papers [43, 44] describe the construction of associative portraits of subject areas, i.e., sets of domain and linguistic knowledge that is the most characteristic to the area. The authors used relation extraction methods that are based on a vector algorithm of determining the words' semantic proximity. The relations were not divided into separate types, but were treated as a single set of associations. The resulting structures were successfully applied to practical problems [45, 46].

It should be noted that most studies do not take the types of relations between words into account. It is difficult to establish whether that is due to the sufficient quality of the solutions to NLP problems that don't consider relation types or the difficulty of defining different types of hierarchical and associative relations. The authors of this study have demonstrated a difference in the degree of influence that synonymic, hierarchical, and associative relations have on using a thesaurus for sentiment analysis [47]. However, the influence that different types of relations between thesaurus terms have on the quality of the results of applying these thesauri to computational linguistic tasks is an understudied problem. Solving

that problem can lead to an improvement and further development of the automatic natural language processing methods.

One method of distinguishing between different types of relations consists of applying lexico-syntactic patterns. Lexico-syntactic patterns are characteristic expressions (phrases and combinations of words), constructions of certain elements of the language. Examples of the patterns that are most common in the Russian language, which include hypo-/hypernymic relations are described rather well in [48].

A team of developers has created a formal lexico-syntactic pattern language (LSPL) [49]. The templates of phrases typical to academic texts, created with the use of the proposed language, were applied by the authors to automatic analysis of scientific and technical Russian-language documents. It should be noted that text processing also involved using traditional dictionaries (terminological and morphological), as well as a glossary of general scientific terms and expressions. However, in this case the templates are used not as a means of extracting relations between domain terms, but as a means of extracting knowledge from NL texts.

Another paper [50] discusses the problem of automatic ontology construction using lexico-syntactic patterns. The author proposed his own lexico-syntactic pattern language and used it as a basis for developing a software package that can be used to store templates and a corpus of Russian-language texts in a database, edit the templates and validate them on the corpus, and conduct semantic analysis of corpus texts. The author has not assessed his method of ontology construction, but he did suggest a way of assessing performance quality based on solving an information retrieval problem.

Most lexico-syntactic patterns are developed by experts. It would be interesting to conduct a study describing and analyzing the wide range of such patterns that are applicable to the task of constructing Russian-language domain-specific thesauri.

To sum up, the methods of automatic extraction of relations between Russian words are mainly used by researchers for constructing lexical structures directly for solving text processing problems. It is important to point out that almost all studies report that using the identified relations results in a significant quality increase. This highlights the need for research into and development of methods of automating the construction of high-quality thesauri, especially in what concerns relation extraction.

An interesting observation is the fact that quality assessment of the linguistic resources constructed in the research discussed in this section is usually based on the quality of the solutions to NLP problems that utilize these resources.

This once again highlights the lack of reference text corpora that has been manually tagged by experts. Since the creation of such corpora is a very time-consuming task, this area is also in need of methods and means of automation.

3. EXPERIMENTS ON RUSSIAN-LANGUAGE THESAURUS CONSTRUCTION

The overview of the existing research on thesaurus construction demonstrates that the creation of specialized linguistic resources has not been sufficiently automatized and, even after automation, requires a lot of input from philology experts on many aspects in order to improve the quality of the thesauri to prepare them for practical application. In order to illustrate the quality of completely automatically constructed thesauri and to show what natural language processing methods are the most useful for generating thesauri, the authors proposed a combined method and conducted experiments that involved creating two domain-specific thesauri.

After analyzing thesaurus construction methods, the authors identified the approaches to keyword and relation extraction that they found the most useful for constructing a thesaurus that can serve as a lexico-semantic model of a domain. This section describes the results of the experiments on the implementation of the developed algorithm. Special attention was paid to the methods of distinguishing different relation types, as that is the most understudied problem.

The Russian-language thesaurus construction algorithm is based on the thesaurus construction algorithm described in the authors' previous study [36] and consists of the following steps:

1. Extracting keywords from texts as thesaurus terms using the TextRank algorithm [51].

2. Extracting associative relations using statistical algorithms.

3. Extracting synonymic relations from existing linguistic resources and by using the Levenshtein distance [53].

4. Extracting hierarchical relations using linguistic methods.

5. Filtering out the terms with no relations.

At each stage of extracting relations, previously extracted relations can be overwritten; for example, if the algorithm has identified a pair of terms as synonyms, then any previously established associative relations between them will be deleted and replaced by the synonymic relation.

After the basic structural elements of the thesaurus have been constructed, the terms for which no relations have been identified are removed, because they cannot be used in further application of the thesaurus for NLP tasks.

3.1. Term Extraction

The first terms to be selected for the thesaurus are the ones that describe the main topics of the subject area, i.e., keywords. Keyword extraction algorithms were considered in the authors' previous study [4]. According to the results of the study, the supervised algorithms are more effective than unsupervised ones, but preparing training material is extremely labor-intensive on the part of the expert linguists. The unsupervised algorithms PageRank and Topical TextRank demonstrate fairly similar performance quality levels. Topical PageRank is a variation of TextRank that more accurately identifies keywords specific for individual texts. Nevertheless, both algorithms output the same final set of keywords of all texts that is formed to be the basis of the thesaurus. Based on the specific features of term extraction methods and the goal to automatize thesaurus construction as much as possible, the algorithm chosen for creating the Russian-language thesaurus was TextRank, which has good quality characteristics and does not require additional training. Experiments on thesaurus construction were carried out on two text bases: articles about Russian migrants selected by expert philologists from the Yaroslavl State University and the corpus of Russian-language tweets (http://linis-crowd.org/).

The corpus of articles on migrants contains 103 texts, of which 20 are positive and 83 are negative. On average, each text contains 682 words or 4785 characters.

The corpus of tweets contains 4320 texts, of which 2160 are positive and 2160 are negative. On average, each text contains 157 words or 1044 characters.

3.2. Associative Relation Extraction

For the task of extracting associations, two algorithms intended for identifying semantically similar terms were selected: latent semantic analysis (LSA) [52] and word2vec (https://code.google.com/archive/p/word2vec/).

The latent semantic analysis method involves constructing a term–document matrix, where the rows correspond to the thesaurus terms extracted at the previous stage and the columns correspond to the domain texts that are used as the basis for generating the thesaurus. The elements of the matrix are the values of the terms' occurrence frequency in specific texts. After the matrix is constructed, it is decomposed by singular values, which results in a matrix of a smaller shape that serves as a fairly accurate approximation of the original matrix. The rows of the final matrix are interpreted as vectors characterizing the relationship of the corresponding term with the texts. These vectors are compared to each other by one of the standard vector proximity measures, for example, by cosine similarity. At the end, the pairs of terms corresponding to vectors with the highest proximity characteristics are assigned associative relations.

The word2vec tool constructs vector representations of words and searches for semantically similar terms by cosine similarity in the same way as LSA. This algorithm requires prior training, but already trained models, including Russian-language ones, exist and can be freely used. Word vectors are constructed using the standard algorithms CBOW (Continuous Bag of Words) and Skip-Gram.

Both algorithms extract a sufficiently large number of qualitative associative relations by statistical methods that do not depend on the features of any one language, so they were chosen for the task of automatic Russian-language thesaurus construction.

3.3. Synonym Extraction

Synonyms were selected by three algorithms that are based on, respectively, the Levenshtein distance [53]; the extraction of synonymic relations from the Synmaster dictionary of synonyms (http://usyn.ru/blog.php?id_blog=11), which contains about 1200000 words and 1 to 20 synonyms to each word; the RuThes thesaurus.

The Levenshtein distance-based method consists of searching for words that have the same root and for wordforms of the same lemma, which are treated as synonyms within the thesaurus. The Levenshtein distance is the number of changes (insertions, deletions, or substitutions of single letters) required to con-

vert one word into the other. This distance and the words' lengths are used to calculate a proximity measure, which will be greater for similar terms than for other pairs.

The algorithms that search for synonyms in existing linguistic resources all work on the same principle: synonymic relations from a dictionary or from the RuThes thesaurus are added to the automatic thesaurus, but only those that connect terms already present in the thesaurus, without adding any new terms.

3.4. Hierarchical Relation Extraction

Hierarchical, or hypo-/hypernymic, relations are selected for the automatic thesaurus by morpho-syntactic linguistic methods [54], or by extracting relations from RuThes. The algorithm for extracting hyponyms and hypernyms from the Ruthes thesaurus is the same as the one for synonyms.

The morpho-syntactic method considers two terms to be connected by hierarchical relations if one of them includes the second as a suffix string or as part of a multiword term. In both cases, the first term is considered to be the hyponym and the second is considered to be the hypernym, since the second one is more general. For example, "customs union" is a hyponym of "union."

3.5. Comprehensive Assessment of the Thesaurus

Metrics for assessing thesaurus quality were taken from the authors' previous study [36], which proposes a comprehensive assessment of a thesaurus that was constructed completely automatically by hybrid methods. Specifically, the chosen graph characteristics were the number of terms, the number of semantic relations of different types and the number of connected components of the thesaurus graph and the size of the largest component.

This comprehensive assessment was chosen for evaluating thesaurus quality because it had been developed by the authors specifically for automatically generated thesauri. Its main advantage is its capability of evaluating the thesaurus on several aspects at once: the quality of terms and semantic relations, structure, and connectivity. The second advantage of the comprehensive assessment is the fact that its calculation is automatic: graph characteristics are calculated completely automatically, do not depend on an expert or third-party resources, and qualitative characteristics are also calculated automatically, with the only additional resource being the alternative thesaurus that is to be used as reference and comparand.

Unfortunately, the standard measures of accuracy and completeness of term and relation extraction cannot be calculated automatically for the proposed thesauri, because for the selected domains there are no thesauri in open access that could be used as comparands. Therefore, the accuracy of term and relation extraction for the thesaurus was assessed manually by an expert.

3.6. Results of the Experiments

A software project that implements the described above method of Russian-language thesaurus construction was created for the experiments on methods of extracting thesaurus relations. The project was implemented and the assessment characteristics of the thesaurus were calculated with the use of the programming languages Python and Java and the libraries NLTK and Gensim, which implement standard methods of natural language data processing.

Table 2 presents a graph-based quality assessment of the thesaurus for the corpus of articles on migrants that was described in subsection 1. It is obvious that the combined method proposed by the authors presents the best thesaurus characteristics: it contains the most terms and relations of all types and also ensures connectivity.

The morpho-syntactic method and the Levenshtein distance method contribute the least: they extract the smallest amount of relations. The largest number of links is extracted by the method based on extracting information from the Synmaster dictionary: using it turns the thesaurus into a single connected component. Using the word2vec, Levenshtein distance, and morpho-syntactic methods also results in a connected thesaurus, but one that is significantly smaller than the others and contains only 728 terms.

For associative relations, the LSA method performs significantly better than word2vec: it extracts about 29 thousand relations, some of which are replaced by synonymic relations at the following algorithm stages, while word2vec extracts about 5 thousand associations, which is about 6 times less.

The number of hierarchical relations in the texts turned out to be very small in comparison with other types of relations, and most of them were extracted from the general purpose thesaurus RuThes.

Table 2. Characteristics of automatically constructed thesauri for the corpus of articles on migrants

Extraction methods for			Quantity of extracted				Quantity of	
hypernyms	synonyms	associations	terms	hypernyms	synonyms	associations	connected comp.	vertices in max. comp.
RuThes	RuThes	LSA	2321	239	484	29345	5	2313
RuThes	Synmaster	LSA	2413	239	59844	27335	1	2413
Morph	Lev	word2vec	728	20	50	5076	1	728
Hybrid	Hybrid	Hybrid	2413	248	60328	32360	1	2413

LSA—latent semantic analysis.
RuThes—the method based on using RuThes.
Synmaster—the method based on using Synmaster.
Lev—the method based on using the Levenshtein distance.
Morph—the morpho-syntactic method.
Hybrid—joint application of the LSA and word2vec methods for associations, the RuThes, Synmaster and Lev methods for synonyms, the Morph and Hybrid methods for hypo- and hypernyms.

Table 3. Characteristics of automatically constructed thesauri for the corpus of tweets

Extraction methods for			Quantity of extracted				Quantity of	
hypernyms	synonyms	associations	terms	hypernyms	synonyms	associations	connected comp.	vertices in max. comp.
RuThes	RuThes	LSA	15629	2808	608	105998	333	14605
RuThes	Synmaster	LSA	15629	2808	1487526	96436	46	15527
Morph	Lev	word2vec	527	80	445	78	49	41
Hybrid	Hybrid	Hybrid	15629	2888	1488577	96476	46	15527

LSA—latent semantic analysis.
RuThes—the method based on using RuThes.
Synmaster—the method based on using Synmaster.
Lev—the method based on using the Levenshtein distance.
Morph—the morpho-syntactic method.
Hybrid—joint application of the LSA and word2vec methods for associations, the RuThes, Synmaster and Lev methods for synonyms, the Morph and Hybrid methods for hypo- and hypernyms.

The experiments with the corpus of tweets (Table 3) repeat the same patterns: the combined method is superior to other methods when they're used separately. The best method for extracting hierarchical relations is the one that uses RuThes; for synonymic relations, the one that uses Synmaster, for associative relations, LSA. The result demonstrated by the statistical method word2vec is worse than for the previous texts sample: it extracts significantly fewer relations, only 78.

Thus, the proposed method demonstrates the best characteristics of the resulting thesaurus's quality. It should be noted that the quality of the result is achieved primarily by the linguistic methods that extract relations from existing linguistic resources. Among the statistical methods, the most effective one was LSA, which extracted a sufficiently large number of relations for both corpora. Other statistical methods either extract only a small number of synonyms (the Levenshtein distance method) or have unstable results that differ from sample to sample (word2vec).

The expert evaluation of the accuracy of term and relation extraction was carried out for the automatic thesaurus constructed using the combined method for the corpus of articles on migrants, as it is much smaller than the other one and can be evaluated by an expert in a reasonable amount of time. The evaluation showed an accuracy of term extraction of 94.3% with only 137 incorrect terms; an accuracy of synonym extraction of 99.9% and hypo- and hypernyms of 98.3%, which is explained by the fact that almost all of them were extracted from credible lexical resources.

As for the structure of the automatically generated thesaurus, it primarily consists of synonyms and associations. The number of hyponyms and hypernyms in it is significantly smaller for both subject areas. A possible reason for this may be the specifics of the considered text corpora, which contain quite a small number terms that are in a hierarchical relationships with each other.

CONCLUSIONS

Analysis of existing thesauri revealed two available resources: RuThes and YARN, both general purpose Russian-language thesauri. Both thesauri can be successfully used for the tasks of automatic text processing. Both are comparable in terms of the number of words and relations between them, although differ in their choice of base units: concepts in the RuThes thesaurus and synsets in YARN.

However, the almost complete absence of available domain thesauri indicates a relevant focus area for thesauri creators. It should also be noted that publishing both lexical resources of this type and models of their development in open access can help improve the quality of thesauri, as well as make a significant contribution to the solutions of natural language processing problems in the relevant areas.

The methods of constructing Russian-language thesauri also need to be developed and analyzed. In that regard, researchers' attention could be directed towards the use of machine learning methods and combining statistical and linguistic methods.

A separate big problem is quality assessment of thesaurus construction methods. The authors distinguish two directions in the development of these methods. The first direction is evaluating the thesaurus itself by assessing both quantitative characteristics of the words and relations contained in the thesaurus and its internal structure. The second direction consists of evaluating the thesaurus indirectly through evaluating the quality of the solutions to natural language processing problems that it was used in. Implementing this approach requires development of open, tagged text corpora.

The analysis of modern studies in the area of Russian-language thesaurus construction and the authors' own experience create a hope that the described problems will be successfully solved in the foreseeable future.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Aitchison, J., Gilchrist, A., and Bawden, D., *Thesaurus Construction and Use: A Practical Manual,* Psychology Press, 2000.
2. Sidorova, E.A., Ontology-based approach to modeling the process of extracting information from text, *Ontol. Proekt.,* 2018, vol. 8, no. 1, pp. 134–151.
3. Elenevskaya, M.N. and Ovchinnikova, I.G., The storage and description of the verbal associations, *Vopr. Psikholingvist.,* 2016, no. 29, pp. 69–92.
4. Paramonov, I., et al., Thesaurus-based method of increasing text-via-keyphrase graph connectivity during keyphrase extraction for e-tourism applications, *Commun. Comput. Inf. Sci.,* 2016, vol. 649, pp. 129–141.
5. Shchitov, I., Lagutina, K., Lagutina, N., and Paramonov, I., Sentiment classification of long newspaper articles based on automatically generated thesaurus with various semantic relationships, *Proceedings of the 21st Conference of Open Innovations Association FRUCT,* Helsinki, 2017, pp. 290–295.
6. Blenda, N. A., Overview of Russian-language thesauri to solve the problem of calculating the semantic similarity for scientific publications, *Informatsionnye tekhnologii i sistemy, Trudy Chetvertoi Mezhdunarodnoi nauchnoi konferentsii* (Information Technologies and Systems, Proceedings of the Fourth International Scientific Conference), 2015, pp. 70–74.
7. Porshnev, S.V., On the quality of open electronic thesauruses of the Russian language, *Sbornik materialov Vserossiiskoi molodezhnoi shkoly-seminara "Aktual'nye problemy informatsionnykh tekhnologii, elektroniki i radiotekhniki—2015 (IT-ER—2015)* (Proc. All-Russian Youth School-Seminar Current Problems of Information Technology, Electronics, and Radio Engineering—2015 (IT-ER—2015), 2015, vol. 2, pp. 45–48.
8. Loukachevitch, N. and Dobrov, B., RuThes linguistic ontology vs. Russian wordnets, *Proceedings of the Seventh Global WordNet Conference,* 2014, pp. 154–162.
9. Loukachevitch, N., Dobrov, B., and Chetviorkin, I., RuThes-Lite, a publicly available version of Thesaurus of Russian language RuThes, *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference Dialogue,* 2014, no. 13, pp. 340–349.
10. Loukachevitch, N.V., Lashevich, G., Gerasimova, A.A., Ivanov, V.V., and Dobrov, B.V., Creating Russian WordNet by conversion, *Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference Dialogue,* 2016, no. 15, pp. 405–415.
11. Braslavski, P., Ustalov, D., Mukhin, M., and Kiselev, Y., YARN: Spinning-in-Progress, *Proceedings of the Eight Global Wordnet Conference,* 2016, pp. 58–65.
12. Sukhonogov, A.M. and Yablonsky, S.A., Automation of the construction of English-Russian WordNet, *Komp'yuternaya lingvistika i intellektual'nye tekhnologii, Trudy Mezhdunarodnogo seminara Dialog* (Computa-

tional Linguistics and Intellectual Technologies. Proceedings of the International Seminar Dialogue), 2005, pp. 25–31.

13. Azarowa, I., RussNet as a computer lexicon for Russian, *Proceedings of the Intelligent Information Systems IIS-2008,* 2008, pp. 341–350.
14. Azarova, I.V., Zakharov, V.P., Kiselev, Yu., Ustalov, D.A., and Khokhlova, M.V., Integration of RussNet and YARN thesauruses, *Komp'yuternaya lingvistika i vychislitel'nye ontologii, Trudy XIX Mezhdunarodnoi obedinennoi nauchnoi konferentsii Internet i sovremennoe obshchestvo (IMS-2016)* (Computational Linguistics and Computational Ontologies, Proceedings of the 19th International United Scientific Conference The Internet and Modern Society (IMS-2016)), St. Petersburg, 2016, pp. 7–13.
15. Sladkova, O., Pirumova, L., and Pirumov A., Internet information resources for agricultural specialists, *Mezhdunar. S-kh. Zh.,* 2016, no. 2, pp. 44–48.
16. Galieva, A.M. and Yakubova, D.D., Principles of representing vocabulary in the socio-political thesaurus of the Tatar language, *Filol. Nauki, Vopr. Teor. Prakt.,* 2016, no. 12-2, pp. 80–84.
17. Galieva, A.M., Kirillovich, A.V., Lukashevich, N.V., Nevzorova, O.A., Suleimanov, D.Sh., and Yakubova, D.D., Russian-tatar socio-political thesaurus: publishing in the linguistic linked open data cloud, *Int. J. Open Inf. Technol.,* 2017, vol. 5, no. 11, pp. 64–73.
18. Ageev, M.S., Dobrov, B.V., and Lukashevich, N.V., Automatic rubrication of texts: Methods and problems, *Uch. Zap. Kazan. Gos. Univ., Ser. Fiz.-Mat. Nauki,* 2008, vol. 150, no. 4, pp. 25–40.
19. Lukashevich, N.V., Dobrov, B.V., Pavlov, A.M., and Shternov, S.V., Ontological resources and information-analytical system in the subject area Security, *Ontol. Proekt.,* 2018, vol. 8, no. 1, pp. 74–95.
20. Mishunin, O.B., Savinov, A.P., and Firstov, D.I., Problems of automatic free-text answer grading in intelligent tutoring systems, *Sovrem. Probl. Nauki Obraz.,* 2015, no. 2-2, pp. 189–199.
21. Alekseev, A.A., Thematic representation of a news cluster as a basis for summarization, *Program. Inzh.,* 2014, no. 3, pp. 41–48.
22. Ustalov, D.A., Concept discovery from synonymy graphs, *Vychisl. Tekhnol.,* 2017, vol. 22, no. S1, pp. 99–112.
23. Kolchin, M., Chistyakov, A., Lapaev, M., and Khaydarova, R., FOODpedia: Russian food products as a linked data dataset, *International Semantic Web Conference,* 2015, pp. 87–09.
24. Hasan, K. and Vincent, N., Automatic keyphrase extraction: A survey of the state of the art, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,* 2014, pp. 1262–1273.
25. Dobrov, B.V. and Lukashevich, N.V., Linguistic ontology on natural sciences and technologies for information-retrieval applications, *Uch. Zap. Kazan. Gos. Univ., Ser. Fiz.-Mat. Nauki,* 2007, vol. 149, no. 2, pp. 49–72.
26. Lukashevich, N.V., Dobrov, B.V., and Chuiko, D.S., Automated analysis of multiword expressions for computational dictionaries, *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Tr. Mezhdunarodnoi konferentsii Dialog* (Computational Linguistics and Intellectual Technologies: Proc. Annual International Conference Dialogue), 2008, no. 7, pp. 339–344.
27. Turney, P.D. and Pantel, P., From frequency to meaning: Vector space models of semantics, *J. Artif. Intell. Res.,* 2010, vol. 37, pp. 141–188.
28. Zakharov, V.P., Corpus-based approach to thesaurus and ontology construction, *Strukt. Prikl. Lingvist.,* 2015, no. 11, pp. 123–141.
29. Kotova, E.E. and Pisarev, I.A., Construction of thematic ontologies using the method of automated thesauri development, *Izv. S.-Peterb. Gos. Electrotekh. Univ. LETI,* 2016, no. 3, pp. 37–47.
30. Ayusheeva, N.N. and Kusheeva, T.N., Method for calculating weight factors of vertices of a semantic network of a scientific text, *Fundam. Issled.,* 2012, no. 6-3, pp. 626–630.
31. Ayusheeva, N.N., Gombozhapova, T.N., and Dorzhaev, T.V., A method for automatically determining the subject of a scientific text, *Fundam. Issled.,* 2016, nos. 8-2, pp. 229–233.
32. Chetviorkin, I. and Loukachevitch, N., Extraction of Russian sentiment lexicon for product meta-domain, *Proceedings of COLING 2012,* 2012, pp. 593–610.
33. Loukachevitch, N. and Levchik, A., Creating a general Russian sentiment lexicon, *Proceedings of Language Resources and Evaluation Conference,* 2016, pp. 1171–1176.
34. Vanyushkin, A.S. and Grashchenko, L.A., Evaluation of keyword extraction algorithms: Tools and resources, *Nov. Inf. Tekhnol. Avtom. Sist.,* 2017, vol. 20, pp. 95–102.
35. Lukashevich, N.V. and Logachev, Yu.M., Automatic term extraction based on feature combination, *Vychisl. Metody Program.,* 2010, vol. 11, no. 4, pp. 108–116.
36. Lagutina, N.S., Lagutina, K.V., Mamedov, E.I., and Paramonov, I.V., Methodological aspects of separating semantic relationships for automatic generation of specialized thesauri and their evaluation, *Model. Anal. Inf. Sist.,* 2016, vol. 23, no. 6, pp. 826–840.
37. Lukashevich, N.V., Quasi-synonyms in linguistic ontologies, *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii "Dialog"* (Computational Linguistics and Intellectual Technologies: Based on the Materials of the Annual International Conference Dialogue), 2010, no. 9, pp. 307–312.

38. Lukashevich, N.V., Modeling of the PART-WHOLE relations in a linguistic resource for information-retrieval applications, *Inf. Tekhnol.,* 2007, no. 12, pp. 28–34.
39. Baranyuk, V.V., Bogoradnikova, A.V., and Smirnova, O.S., Defining the scope semantics by forming its thesaurus, *Int. J. Open Inf. Technol.,* 2016, vol. 4, no. 9, pp. 74–79.
40. Nugumanova, A.B., Bessmertnyi, I.A., Petsina, P., and Baiburin, E.M., Semantic relations in text classification based on bag-of-words model, *Program. Prod. Sist.,* 2016, no. 2, pp. 89–99.
41. Panchenko, A., Ustalov, D., Arefyev, N., Paperno, D., Konstantinova, N., Loukachevitch, N., and Biemann, C., Human and machine judgements for Russian semantic relatedness, *Analysis of Images, Social Networks and Texts. 5th International Conference, AIST 2016,* 2016, pp. 221–235.
42. Rapp, R., The automatic generation of thesauri of related words for English, French, German, and Russian, *Int. J. Speech Technol.,* 2008, vol. 11, nos. 3–4, pp. 147–156.
43. Galina, I.V., Kozerenko, E.B., Morozova, Yu.I., Somin, N.V., and Sharnin, M.M., Associative portraits of subject areas as a tool for automated construction of big data systems for knowledge extraction: Theory, methods, visualization, and application, *Inf. Primen.,* 2015, vol. 9, no. 2, pp. 92–110.
44. Kuznetsov, I.P., Kozerenko, E.B., and Charnine, M.M., Technological peculiarity of knowledge extraction for logical-analytical systems, *Proceedings of ICAI,* 2012, vol. 12, pp. 18–21.
45. Zolotarev, O.V. and Sharnin, M.M., Methods for extracting knowledge from natural language texts and the construction of models of business processes on the basis of identifying processes, objects, their relationships, and characteristics, *Trudy Mezhdunarodnoi nauchnoi konferentsii CPT2014* (Proceedings of the International Scientific Conference CPT2014), 2015, pp. 92–98.
46. Zolotarev, O.V., Sharnin, M.M., and Klimenko, S.V., Semantic approach to the analysis of terrorist activity on the Internet based on thematic modeling methods, *Vestn. Ross. Nov. Univ., Ser.: Slozhnye Sist.: Modeli Anal. Upr.,* 2016, no. 3, pp. 64–71.
47. Lagutina, N.S, Lagutina, K.V., Shchitov, I.A., and Paramonov, I.V., Analysis of influence of different relations types on the quality of thesaurus application to text classification problems, *Model. Anal. Inf. Sist.,* 2017, vol. 24, no. 6, pp. 772–787.
48. Sabirova, K. and Lukanin, A., Automatic extraction of hypernyms and hyponyms from Russian texts, *Supplementary Proceedings of the 3rd International Conference on Analysis of Images, Social Networks and Texts (AIST'2014),* 2014, pp. 35–40.
49. Bolshakova, E.I., Ivanov, K.M., Sapin, A.S., and Sharikov, G.F., A system for extracting information from texts on the basis of lexical and syntactic templates, *Pyatnadtsataya natsional'naya konferentsiya po iskusstvennomu intellektu s mezhdunarodnym uchastiem* (Fifteenth National Conference on Artificial Intelligence with International Participation), 2016, pp. 14–22.
50. Rabchevskii, E.A., Automatic construction of ontologies based on lexical and syntactic patterns for information retrieval, *Elektronnye biblioteki: Perspektivnye metody i tekhnologii, elektronnye kollektsii, Sb. nauch. tr. 11-i Vserossiiskoi nauchnoi konferentsii RCDL-2009* (Digital Libraries: Promising Methods and Technologies, Digital Collections, Proc. 11th All-Russian Scientific Conference RCDL-2009), Petrozavodsk, 2009, pp. 69–77.
51. Mihalcea, R. and Tarau, P., TextRank: Bringing order into texts, *Proceedings of Empirical Methods in Natural Language Processing—EMNLP,* Barcelona, 2004, pp. 404–411.
52. Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A., Latent semantic analysis, *Proceedings of the 16th International Joint Conference on Artificial Intelligence,* 2004, pp. 1–14.
53. Noh, S., Kim, S., and Jung, C., A lightweight program similarity detection model using XML and Levenshtein distance, *FECS,* 2006, pp. 3–9.
54. Lefever, E., Van de Kauter, M., and Hoste, V., Evaluation of automatic hypernym extraction from technical corpora in English and Dutch, *9th International Conference on Language Resources and Evaluation (LREC),* 2014, pp. 490–497.

Translated by A. Ovchinnikova