Ontologies for Natural Language Processing: Case of Russian

Natalia Loukachevitch
Research Computing Center of
Lomonosov Moscow State University
louk_nat@mail.ru

Boris Dobrov
Research Computing Center of
Lomonosov Moscow State University
dobrov_bv@mail.ru

Abstract

The paper describes the RuThes family of Russian thesauri intended for natural language processing and information retrieval applications. RuThes-like thesauri include, besides RuThes, Sociopolitical thesaurus, Security Thesaurus, and Ontology on Natural Sciences and Technologies. The RuThes format is based on three approaches for developing computer resources: Princeton WordNet, information-retrieval thesauri, and formal ontologies. The published version of RuThes thesaurus (RuThes-lite 2.0) became a basis for semi-automatic generation of RuWordNet, WordNet-like thesaurus for Russian. Currently researchers can use both RuThes-lite or RuWordNet and compare them in applications. Other RuThes-like resources are being prepared to publication.

1. Introduction

Term ”ontology” is used in broader and narrower senses. In a broader sense, ontology is considered as an umbrella concept for various resources such as glossaries, thesauri, taxonomies, subject headings, or formal axiomatic systems. This understanding corresponds to the well-known definition of an ontology as ”a formal specification of a shared conceptualization” (Guarino et al., 2009), (Gruber, 1995), (Studer et al., 1998), because all these resources represent some conceptualization of the external world.

In a narrower sense, ontology is a formal representation system of concepts defined using logical formalism to explain the meanings in a computable way (Pease, 2011). Such ontologies should be independent of any specific natural language (Nirenburg and Raskin, 2004), (Nirenburg and Wilks, 2001). The main author of well-known CYC ontology Doug Lenat wrote that taking the meanings of words into account can only confuse, the meanings of words divide the world ambiguously, and the division lines come from a variety of reasons: historical, physiological, etc. (Lenat et al., 1995).

In contrast to the above-mentioned approaches, the WordNet thesaurus is often mentioned as a linguistic, or lexical ontology, that is an ontology, whose concepts are mainly based on senses of existing lexical units, the terms of the subject field (Magnini and Speranza, 2002), (Veale and Hao, 2008). Linguistic ontologies cover most of the words of the language or a subject field, and at the same time they have an ontological structure represented in relations between the concepts. Synsets of WordNet are often considered as lexicalized concepts. Later, the WordNet structure was reproduced in various WordNet-like resources (wordnets) created for many languages (Azarowa, 2008), (Derwojedowa et al., 2008), (Koeva, 2010), (Kunze and Lemnitzer, 2010).

However, WordNet has been created as a lexical rather than ontological resource (Fellbaum, 1998)). in the framework of relational semantics (Miller et al., 1990). WordNet is mainly intended to describe lexical relations, which is quite different from the primary aim of ontologies to describe knowledge about the world, not about language. The WordNet structure was criticized from the ontological point of view (Guarino, 1998). Guarino and Welty (Guarino and Welty, 2009), developed the OntoClean approach for stricter description of relations and applied it to WordNet. Other authors (Wilks, 2009) suppose that NLP resources such as WordNet should not be subjects of such strict procedures, because of vagueness of their units, word senses.

Conventional information retrieval thesauri can also be considered as linguistic ontologies because they are based on real terms of a subject field (Will, 2012), (Clarke and Zeng, 2012). A term is defined as one or more words referring to a concept; a concept is considered as a unit of thought, regardless of the terms that express it (NISO, 2005). Contemporary standards for developing of information-retrieval thesauri stress that thesaurus relations are established between concepts, not between terms (Clarke and Zeng, 2012). However, information-retrieval thesauri are not intended for use in automatic processing of texts: they should be used in manual indexing by human experts for improvement of information retrieval in physical or digital libraries.

Thus, there exist different approaches to representing models of linguistic ontologies for natural language processing on the scale from more lexical to more conceptual resources. In this paper, we consider the approach to developing Russian ontological resources having the format of the RuThes thesaurus (Loukachevitch and Dobrov, 2014) and created for automatic processing of documents in information-analytical systems and natural language processing. These resources are linguistic ontologies uniting some principles of their organization from WordNet, information-retrieval thesauri and formal ontologies. They were utilized in various information-retrieval and NLP applications (Loukachevitch and Dobrov, 2014). RuThes was successfully evaluated in text summarization (Mani et al., 2002), text clustering (Loukachevitch et al., 2017), text categorization (Loukachevitch and Dobrov, 2014), detecting Russian paraphrases (Loukachevitch et al., 2017), etc.

We compare the RuThes model with the WordNet-like model of knowledge representation and describe some applications of RuThes-like resources for text analytics. We also consider the structure and current state of RuWordNet, WordNet-like Russian thesaurus, semi-automatically generated from RuThes data. The specificity of RuWordNet generation allows better understanding of the differences between representation models of the thesauri.

2. Existing Russian Thesauri

For the Russian language, there were at least four known projects for creating a wordnet. In the RussNet project (Azarowa, 2008), the authors planned to create a Russian wordnet from scratch, guided by the principles of Princeton WordNet. In two different projects (Gelfenbeyn et al., 2003), (Balkova et al., 2008), attempts were made to automatically translate WordNet into Russian, with all the original thesaurus structure preserved. The results of (Gelfenbeyn et al., 2003) have been published[1], but the analysis of the thesaurus generated in this way shows that it requires considerable editing efforts.

The last Russian wordnet project YARN (Yet Another Russian wordNet) (Braslavski et al., 2016) was initiated in 2012 and initially was created on the basis of crowdsourcing, i.e. involvement of a large number of non-professional native speakers. Currently, YARN contains a significant number of synsets with a small number of relationships between them. The published version of the YARN[2] thesaurus contains too many similar or partially similar synsets introduced by different participants.

In (Azarova et al., 2016), the authors describe a current project on the integration of the RussNet thesaurus (Azarowa, 2008) and the YARN thesaurus YARN (Braslavski et al., 2016) into a single linguistic resource, where the expert approach and the crowdsourcing will be combined.

For Russian, traditional information-retrieval thesauri in social sciences and the humanities have been developed and are supported in the Institute of Scientific Information of Russian Academy of Sciences (INION RAN). This institution publishes separate issues of thesauri on economics, sociology, linguistics etc., which were developed according to the guidelines of international and national standards. These thesauri cannot be used for automatic processing of document and news flows because they are, in fact, lists of selected keywords, denoting the most significant concepts of the domain, with low coverage of real texts (Mdivani, 2013). There are also several Russian versions of international information-retrieval thesauri or controlled vocabularies (Lipscomb, 2000), (Kupriyanov et al., 2016).

[1] wordnet.ru

[2] https://russianword.net/

3. RuThes Family of Resources

The structure of the RuThes thesaurus of the Russian language is based on three approaches for developing computer resources: information-retrieval thesauri, WordNet-like thesauri, and formal ontologies.

The RuThes thesaurus is created in form of a linguistic ontology, which concepts are based on senses of really existing words and phrases. RuThes is a concept-oriented resource as much as possible in describing senses of Russian words and expressions. Each concept has a unique, unambiguous name. In this, RuThes is similar to information-retrieval thesauri and formal ontologies. Rules for inclusion of phrases in the thesaurus are more similar to information-retrieval thesauri guidelines (NISO, 2005).

Each concept is linked with words and phrases conveying the concept in texts (text entries). Detailed description of lexical units (words in specific senses), representation of senses of ambiguous words are closer to wordnets. Types of relations between concepts originate from information-retrieval thesauri, but some explications are made on the basis of ontological studies. There exist several large Russian thesauri presented in the same format:

- RuThes thesaurus comprising words and phrases of literary Russian together with terms of so-called sociopolitical domain (see below) (Loukachevitch and Dobrov, 2014);
- RuThes-lite[3], a published version of RuThes, can be obtained for non-commercial purposes (Loukachevitch et al., 2014);
- Sociopolitical Thesaurus comprising lexical items and terms from the sociopolitical domain. The sociopolitical domain is a broad domain describing everyday life of modern society and uniting many professionals domains, such as politics, law, economy, international relations, finances, military affairs, arts, and others. Terms of this domain are usually known to not only professionals, but also to ordinary people (Loukachevitch and Dobrov, 2015). Thus, this thesaurus contains important knowledge for processing news flow, legal documents, and developing new domain-specific resources. The Sociopolitical thesaurus can exist and be used separately. At the same time it is included as a part into three larger thesauri: RuThes, OENT ontology, and the Security Thesaurus;
- Ontology on Natural Sciences and Technologies (OENT) includes terms of mathematics, physics, chemistry, geology, astronomy etc., terms of technological domains (oil and gas, power stations, cosmic technologies, aircrafts, etc.). It also contains the Sociopolitical thesaurus as a part because scientific and technological problems can be discussed together with political, economical, industrial, and other issues (Dobrov and Loukachevitch, 2006);
- Security Thesaurus is an extension of the RuThes thesaurus and includes terminology related to social, national and religious conflicts, extremism and terrorism, information security.

The Table 1 contains quantitative characteristics of the above-mentioned resources.

Table 1: RuThes-like Thesauri

Thesaurus	Number of concepts	Number of Text Entries	Number of Conceptual Relations
RuThes	55,275	170,130	226,743
RuThes-lite	31,540	111,559	128,866
Sociopolitical Thesaurus	41,426	121,292	161,523
OENT	94,103	262,955	376,223
Security Thesaurus	66,810	236,321	271,297

[3] www.labinform.ru/pub/ruthes/index.htm

4. Specific Features of RuThes Structure

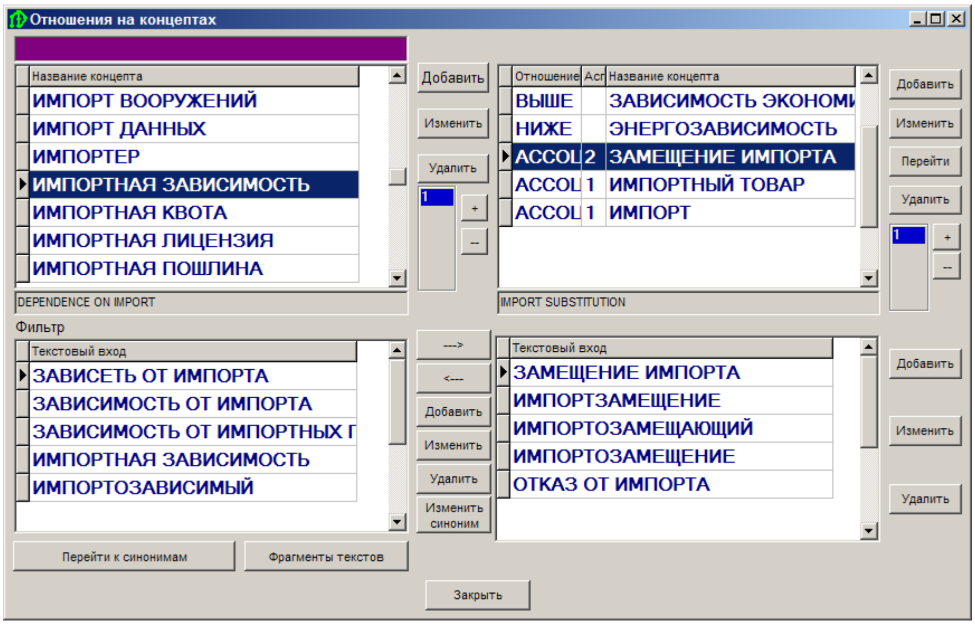
The main unit of RuThes is a concept as a unit of thought regardless of words expressing it. A concept has a unique, unambiguous name. Concept names are similar to descriptors in information-retrieval thesauri, that is precisely formulated terms referring to implied concepts. If an unambiguous and clear name in form of an existing word or a phrase cannot be found, than an ambiguous word can be used for naming and supplied with a “relator” (a brief note in parentheses).

The RuThes concepts are not divided into part-of-speech-oriented nets as in wordnets. This approach is closer to formal ontologies. Therefore, text entries of a specific concept can comprise single words of different parts of speech, including ambiguous ones, and phrases that can be either idiomatic or compositional groups. Large rows of synonyms and term variants are collected to provide better recognition of concepts in texts. The concept-based approach seems to be more convenient for text analytics and information-analytical systems in specific domains.

Fig. 1-2 show the interface of thesaurus developing. The upper left form contains a list of concepts in alphabetical order. Fig. 1 shows concepts from the Sociopolitical thesaurus: *Import of weapons, Import of information, Importer, Import dependence, Import quota, Import license, Import tax*. The lower left form shows text entries for the highlighted concept (*Import dependence*), which include: *to depend on import, import dependence, import-dependent, dependence on imported goods, etc.*

The right upper form presents the relations of the highlighted concept. Fig. 1 shows the relation of *Import dependence* concept with such concepts as *Economy dependence, Energy dependence, Import substitution, Imported goods,* and *Import*. The lower right form shows text entries for a related concept. The low right form of Fig. 1 describes text entries of *Import substitution* concept.

Fig. 2 shows a fragment from the Security thesaurus. The visible list of concepts includes: *Attack on payment system, Attack on search engine, Attack on embassy, Attack on vulnerability, Attack on torrent, Zero-day attack, DOS-attack*. Text entries of *Attack on vulnerability* concept comprise in particular such a phrase as *exploiting vulnerability*. It should be noted that this is a true, but non-evident synonym of *attack on vulnerability*, found in real texts.



Отношения на концептах

Название концепта
ИМПОРТ ВООРУЖЕНИЙ
ИМПОРТ ДАННЫХ
ИМПОРТЕР
ИМПОРТНАЯ ЗАВИСИМОСТЬ
ИМПОРТНАЯ КВОТА
ИМПОРТНАЯ ЛИЦЕНЗИЯ
ИМПОРТНАЯ ПОШЛИНА

DEPENDENCE ON IMPORT

Отношение	Аст	Название концепта
ВЫШЕ		ЗАВИСИМОСТЬ ЭКОНОМИ
НИЖЕ		ЭНЕРГОЗАВИСИМОСТЬ
АССОЦ	2	ЗАМЕЩЕНИЕ ИМПОРТА
АССОЦ	1	ИМПОРТНЫЙ ТОВАР
АССОЦ	1	ИМПОРТ

IMPORT SUBSTITUTION

Фильтр

Текстовый вход
ЗАВИСЕТЬ ОТ ИМПОРТА
ЗАВИСИМОСТЬ ОТ ИМПОРТА
ЗАВИСИМОСТЬ ОТ ИМПОРТНЫХ Г
ИМПОРТНАЯ ЗАВИСИМОСТЬ
ИМПОРТОЗАВИСИМЫЙ

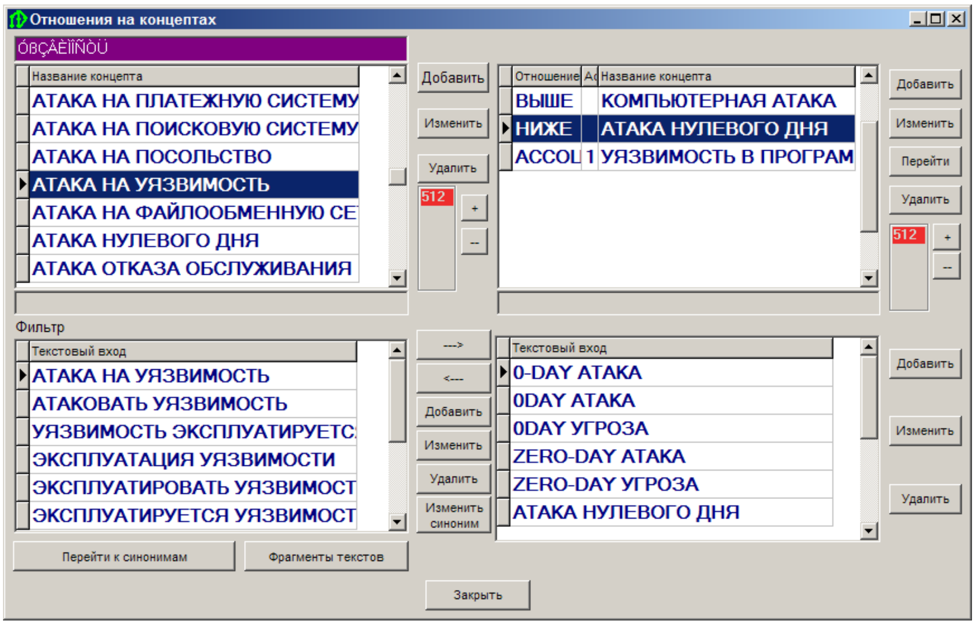
Текстовый вход
ЗАМЕЩЕНИЕ ИМПОРТА
ИМПОРТЗАМЕЩЕНИЕ
ИМПОРТОЗАМЕЩАЮЩИЙ
ИМПОРТОЗАМЕЩЕНИЕ
ОТКАЗ ОТ ИМПОРТА

Figure 1: Relations and text entries for *import dependence* concept from Sociopolitical thesaurus

In RuThes-lite, thesauri there are four basic types of relationships between concepts. The first type of the relations is the class-subclass relationship, it has the properties of transitivity and inheritance.

The second type of the RuThes relations is the part-whole relation. An important condition for establishing this relationship in RuThes is that the concept-parts must be rigidly connected with their whole, that is, each example of the concept-part must, throughout its entire existence, require the existence of the concept-whole. This corresponds to the guidelines of information-retrieval thesaurus standards recommending that part-whole relations should be established when the part-concept ”inherently included” in the whole-concept, regardless of context (NISO, 2005). This idea can be explicated in ontological terms of inseparable parts or mandatory wholes (Guizzardi, 2011).

Under these conditions, it is possible to rely on the transitivity property of the part-whole relation, which is very important for automatic logical inference in the process of automatic text processing (Loukachevitch and Dobrov, 2015). In RuThes, the part-whole relation is used not only to describe physical parts, but also to other internal attributes, such as properties or roles for situations (Guarino et al., 2009).



Отношения на концептах

ÓßÇÂÈÌÎÑÒÜ

Название концепта
АТАКА НА ПЛАТЕЖНУЮ СИСТЕМУ
АТАКА НА ПОИСКОВУЮ СИСТЕМУ
АТАКА НА ПОСОЛЬСТВО
АТАКА НА УЯЗВИМОСТЬ
АТАКА НА ФАЙЛООБМЕННУЮ СЕ
АТАКА НУЛЕВОГО ДНЯ
АТАКА ОТКАЗА ОБСЛУЖИВАНИЯ

Отношение	Ас	Название концепта
ВЫШЕ		КОМПЬЮТЕРНАЯ АТАКА
НИЖЕ		АТАКА НУЛЕВОГО ДНЯ
АССОЦ	1	УЯЗВИМОСТЬ В ПРОГРАМ

Фильтр

Текстовый вход
АТАКА НА УЯЗВИМОСТЬ
АТАКОВАТЬ УЯЗВИМОСТЬ
УЯЗВИМОСТЬ ЭКСПЛУАТИРУЕТС
ЭКСПЛУАТАЦИЯ УЯЗВИМОСТИ
ЭКСПЛУАТИРОВАТЬ УЯЗВИМОСТ
ЭКСПЛУАТИРУЕТСЯ УЯЗВИМОСТ

Текстовый вход
0-DAY АТАКА
0DAY АТАКА
0DAY УГРОЗА
ZERO-DAY АТАКА
ZERO-DAY УГРОЗА
АТАКА НУЛЕВОГО ДНЯ

Figure 2: Relations and text entries for *zero-day vulnerability* concept from Sociopolitical thesaurus

Two other relations asymmetric and symmetric associations originate from the related term relation in information retrieval thesauri (NISO, 2005).

The asymmetric association $asc_1 - asc_2$, connects two concepts that cannot be linked by the class-subclass or part-whole relationships, but when one of which does not exist without the existence of another, for example, the *import dependence* concept can exist if only the *import* concept exists, or the *vulnerability attack* concept can appear if only the *computer vulnerability* concept exists. This relation is close to the external ontological dependence relation in ontological terms (Guarino et al., 2009).

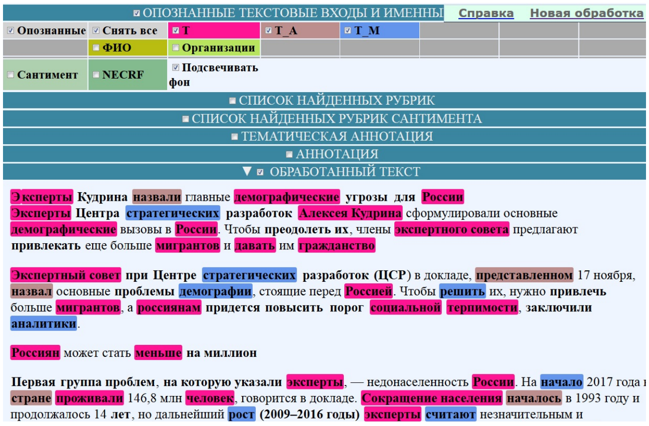
The last type of relationships is the symmetrical association, it links concepts that are very similar in meaning, but which seems difficult to represent as one concept.

Thus, the system of the RuThes thesaurus relations describes the most significant relationships of concepts. It originates from relations used in conventional information-retrieval thesauri and explicated in existing ontological terms.

5. RuThes Ontologies in Information-Analytical Systems

RuThes ontologies are used in information-analytical systems as a tool of conceptual indexing and search, various forms of query expansion can be carried out. One of the main applications of RuThes-like ontologies is text categorization and other tasks of text analytics.

The mainstream technology of automatic document categorization is the machine-learning approach. This approach assumes that there is a sufficient training collection for learning the algorithms. However, many organizations have a need in automatic text categorization, when even a category system (system



☑ ОПОЗНАННЫЕ ТЕКСТОВЫЕ ВХОДЫ И ИМЕННЫ Справка Новая обработка

☑ Опознанные ☑ Снять все ☑ Т ☑ Т_А ☑ Т_М

□ ФИО □ Организации

□ Сантимент □ NECRF ☑ Подсвечивать фон

□ СПИСОК НАЙДЕННЫХ РУБРИК

□ СПИСОК НАЙДЕННЫХ РУБРИК САНТИМЕНТА

□ ТЕМАТИЧЕСКАЯ АННОТАЦИЯ

□ АННОТАЦИЯ

▼ ☑ ОБРАБОТАННЫЙ ТЕКСТ

Эксперты Кудрина назвали главные демографические угрозы для России
Эксперты Центра стратегических разработок Алексея Кудрина сформулировали основные демографические вызовы в России. Чтобы преодолеть их, члены экспертного совета предлагают привлекать еще больше мигрантов и давать им гражданство

Экспертный совет при Центре стратегических разработок (ЦСР) в докладе, представленном 17 ноября, назвал основные проблемы демографии, стоящие перед Россией. Чтобы решить их, нужно привлечь больше мигрантов, а россиянам придется повысить порог социальной терпимости, заключили аналитики.

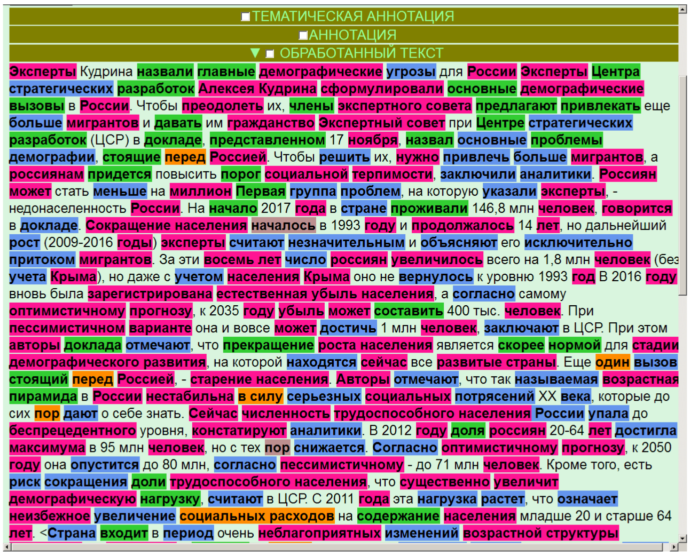
Россиян может стать меньше на миллион

Первая группа проблем, на которую указали эксперты, — недонаселенность России. На начало 2017 года
стране проживали 146,8 млн человек, говорится в докладе. Сокращение населения началось в 1993 году и продолжалось 14 лет, но дальнейший рост (2009–2016 годы) эксперты считают незначительным и

Figure 3: Security thesaurus terms found in a text. Brown and blue boxes show ambiguous terms, which should be disambiguated

of subject headings) may be absent and should be created from scratch or with the use of existing similar categorial systems.

In such conditions, machine-learning approaches cannot be applied, and knowledge-based methods of text categorization, i.e. exploiting manual rules for describing categories, are more acceptable. When one creates a hierarchical system of categories and rules for the text categorization in a broad subject domain, it is convenient to use the thesaurus support, because the thesaurus allows working not with separate words and expressions, but with concepts and substructures of the thesaurus (Loukachevitch and Dobrov, 2015).



□ ТЕМАТИЧЕСКАЯ АННОТАЦИЯ

□ АННОТАЦИЯ

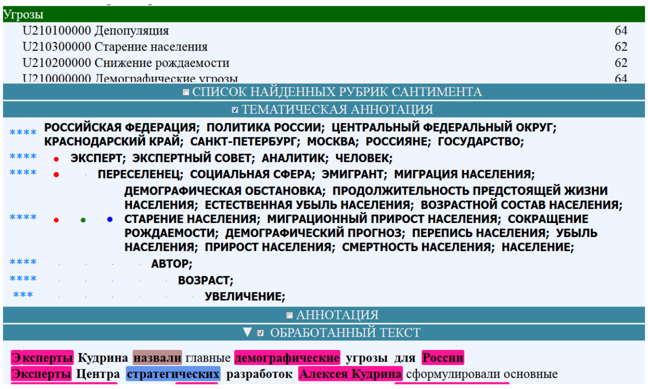
▼ □ ОБРАБОТАННЫЙ ТЕКСТ

Эксперты Кудрина назвали главные демографические угрозы для России Эксперты Центра стратегических разработок Алексея Кудрина сформулировали основные демографические вызовы в России. Чтобы преодолеть их, члены экспертного совета предлагают привлекать еще больше мигрантов и давать им гражданство Экспертный совет при Центре стратегических разработок (ЦСР) в докладе, представленном 17 ноября, назвал основные проблемы демографии, стоящие перед Россией. Чтобы решить их, нужно привлечь больше мигрантов, а россиянам придется повысить порог социальной терпимости, заключили аналитики. Россиян может стать меньше на миллион Первая группа проблем, на которую указали эксперты, - недонаселенность России. На начало 2017 года в стране проживали 146,8 млн человек, говорится в докладе. Сокращение населения началось в 1993 году и продолжалось 14 лет, но дальнейший рост (2009-2016 годы) эксперты считают незначительным и объясняют его исключительно притоком мигрантов. За эти восемь лет число россиян увеличилось всего на 1,8 млн человек (без учета Крыма), но даже с учетом населения Крыма оно не вернулось к уровню 1993 год В 2016 году вновь была зарегистрирована естественная убыль населения, а согласно самому оптимистичному прогнозу, к 2035 году убыль может составить 400 тыс. человек. При пессимистичном варианте она и вовсе может достичь 1 млн человек, заключают в ЦСР. При этом авторы доклада отмечают, что прекращение роста населения является скорее нормой для стадии демографического развития, на которой находятся сейчас все развитые страны. Еще один вызов стоящий перед Россией, - старение населения. Авторы отмечают, что так называемая возрастная пирамида в России нестабильна в силу серьезных социальных потрясений XX века, которые до сих пор дают о себе знать. Сейчас численность трудоспособного населения России упала до беспрецедентного уровня, констатируют аналитики. В 2012 году доля россиян 20-64 лет достигла максимума в 95 млн человек, но с тех пор снижается. Согласно оптимистичному прогнозу, к 2050 году она опустится до 80 млн, согласно пессимистичному - до 71 млн человек. Кроме того, есть риск сокращения доли трудоспособного населения, что существенно увеличит демографическую нагрузку, считают в ЦСР. С 2011 года эта нагрузка растет, что означает неизбежное увеличение социальных расходов на содержание населения младше 20 и старше 64 лет. <Страна входит в период очень неблагоприятных изменений возрастной структуры

Figure 4: RuThes coverage of the same text

In RuThes-based text categorization, each category is represented by the disjunction of alternatives, each clause is a conjunction. The conjuncts, in turn, are described by experts with the help of so-called ”support” concepts. For each support concept, the rule of its extension f (•) is defined, which determines what subordinate concepts should be included in the category profile: without an extension (denoted by the symbol ”N”), the full extension in the hierarchy tree (”Y”), extension only by subclass relations (”L”), etc. In every step, automatically derived concepts in the category description can be edited.

For example, the music category can be described with the single concept *Musical art*$_Y$, where Y means full expansion to lower levels of the hierarchy, including hyponyms, parts, and dependent concepts. The full Boolean expression for this category looks like a disjunction of more than 400 concepts, including musical styles, musical instruments, musicians, musical compositions, musical performances, musical groups and organizations: *Adagio or Accordion or ... Jazz music or . . . Musician ... or Opera or . . . or Orchestra* or . . . (etc.).



Угрозы

U210100000 Депопуляция 64

U210300000 Старение населения 62

U210200000 Снижение рождаемости 62

U210000000 Демографические угрозы 64

□ СПИСОК НАЙДЕННЫХ РУБРИК САНТИМЕНТА

☑ ТЕМАТИЧЕСКАЯ АННОТАЦИЯ

**** РОССИЙСКАЯ ФЕДЕРАЦИЯ; ПОЛИТИКА РОССИИ; ЦЕНТРАЛЬНЫЙ ФЕДЕРАЛЬНЫЙ ОКРУГ; КРАСНОДАРСКИЙ КРАЙ; САНКТ-ПЕТЕРБУРГ; МОСКВА; РОССИЯНЕ; ГОСУДАРСТВО;

**** ● ЭКСПЕРТ; ЭКСПЕРТНЫЙ СОВЕТ; АНАЛИТИК; ЧЕЛОВЕК;

**** ● ПЕРЕСЕЛЕНЕЦ; СОЦИАЛЬНАЯ СФЕРА; ЭМИГРАНТ; МИГРАЦИЯ НАСЕЛЕНИЯ;

**** ● ● ● ДЕМОГРАФИЧЕСКАЯ ОБСТАНОВКА; ПРОДОЛЖИТЕЛЬНОСТЬ ПРЕДСТОЯЩЕЙ ЖИЗНИ НАСЕЛЕНИЯ; ЕСТЕСТВЕННАЯ УБЫЛЬ НАСЕЛЕНИЯ; ВОЗРАСТНОЙ СОСТАВ НАСЕЛЕНИЯ; СТАРЕНИЕ НАСЕЛЕНИЯ; МИГРАЦИОННЫЙ ПРИРОСТ НАСЕЛЕНИЯ; СОКРАЩЕНИЕ РОЖДАЕМОСТИ; ДЕМОГРАФИЧЕСКИЙ ПРОГНОЗ; ПЕРЕПИСЬ НАСЕЛЕНИЯ; УБЫЛЬ НАСЕЛЕНИЯ; ПРИРОСТ НАСЕЛЕНИЯ; СМЕРТНОСТЬ НАСЕЛЕНИЯ; НАСЕЛЕНИЕ;

**** АВТОР;

**** ВОЗРАСТ;

*** УВЕЛИЧЕНИЕ;

□ АННОТАЦИЯ

▼ ☑ ОБРАБОТАННЫЙ ТЕКСТ

Эксперты Кудрина назвали главные демографические угрозы для России

Эксперты Центра стратегических разработок Алексея Кудрина сформулировали основные

Figure 5: The upper part shows the threat categories found for a text. The central part presents topic nodes of related concepts such as ”demographic situation” node.

The automatic text categorization is based on the automatically constructed thematic representation of documents that models the main topic and sub-topics of the document in sets (thematic nodes) of similar concepts mentioned in the document (Loukachevitch and Dobrov, 2015). Such a basis for the text categorization makes it possible to process texts of different types and sizes: normative acts, newspaper articles, news reports, scientific publications, or sociological surveys.

5.1. Automatic Document Processing for Text Analytics

The main stages of thesaurus-based document processing include:

- Tokenization and lemmatization, that is, the transfer of word forms to dictionary forms (lemmas);

- Matching with the thesaurus based on the lemma representation of the document. Multiword terms from a thesaurus are matched with the text using lemma sequences. Fig. 3 shows the term coverage of news text ”Kudrin’s experts named the main demographic threats for Russia”[4]), according to the Security thesaurus. Fig. 4 shows the coverage of matching the same text with RuThes text entries;

- Disambiguation of ambiguous text entries. Brown and blue boxes on Fig. 3 highlight ambiguous terms, which were automatically resolved. For example, Russian word *demografiya (demography)* can mean *demographic situation* or *demographic science*. The quality of the disambiguation procedure was previously evaluated as 75% (F-measure) for domain-specific thesauri (Security thesaurus and Sociopolitical thesaurus). Green and orange boxes on Fig. 4 show ambiguous words that were not disambiguated in the current processing. The quality of disambiguation for RuThes is much lower than for domain-specific thesauri because of the presence of ambiguous general words;

- Grouping semantically related concepts into so-called thematic nodes. This provides better determination of concept weights, which are calculated on the basis of the concept frequency in the given document and the significance of the corresponding thematic node. Fig. 5 (in the center) demonstrate such thematic nodes for the above-mentioned document about the demographic threats. The important thematic node about the demographic situation includes the following concepts: *Demographic situation, Life expectancy, Natural population decline, Age structure of population, Population aging, Net migration rate, Decline in birth rate, Demographic prognosis,* etc.;

- Forming the conceptual index of the document. Conceptual index of a document consist of concepts found in the document and their assigned weights. The weight of a concept accounts for the

[4] (https://www.rbc.ru/economics/17/11/2017/5a0eb1d39a79470f724250b4

significance of the corresponding thematic node and the frequency of the concept in the document. In the example text, the important threat ”population aging” was explicitly mentioned only once in the text, and it could obtain a too low frequency-based weight, but with the support of the main topic node ”demographic situation”, its weight is considerably higher;

- Calculation of category weights in dependence of concepts included into the rules of the inference for this category. Fig. 5 (upper part) shows the categories found in the mentioned document, including ”Depopulation”, ”Population aging”, ”Fertility decline”;
- The results of document processing, including the word index, the conceptual index, the calculated categories, etc. are loaded into an information-analytical system.

6. Generating of Russian WordNet from RuThes-lite

As it was described in Section 2, there did not exist large and qualitative Russian wordnet, but there is a demand from researchers to have a thesaurus in the WordNet format for Russian. Therefore such a thesaurus called RuWordNet was semi-automatically generated from the published version of RuThes (RuThes-lite 2.0) (Loukachevitch et al., 2018).

Table 2: Quantitative characteristics of synsets and entries in RuWordNet

Part of Speech	Number of Synsets	Number of Unique Entries	Number of Senses
Nouns	29,296	68,695	77,153
Verbs	7,634	26,356	35,067
Adjectives	12,864	15,191	18,195

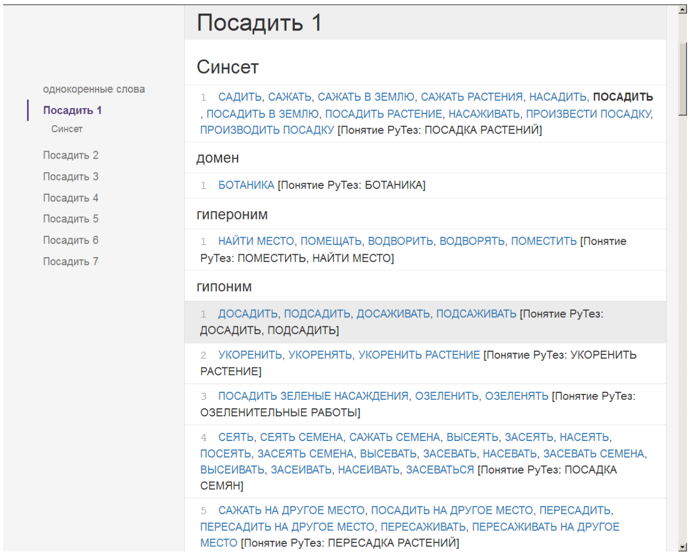
The main above-mentioned differences of RuThes from WordNet-like thesauri are as follows: representation of word senses without division to parts of speech in a single hierarchical net, and conceptual relations. Thus, the first step to construct RuWordNet was to divide the source resource into three nets of nouns, verbs, and adjectives. This subdivision was based on the morpho-syntactic representation of RuThes-lite 2.0 text entries, that is part-of-speech labels for single words and syntactic classes (noun group, verb group, and adjective group) for phrases. The divided synsets were linked to each other with the relation of part-of-speech synonymy (cross-categorial synonymy). The Table 2 contains quantitative characteristics of the RuWordNet synsets, senses, and unique entries.

The hyponym-hypernym relations were established between synsets of the same part of speech. These relations include direct hyponym-hypernym relations from RuThes-lite 2.0. In addition, the transitivity property of hyponym-hypernym relations was employed in cases when a specific synset did not contain a specific part of speech but its parent and child had text entries of this part of speech. In such cases, the hypernymy-hyponymy relation was established between the child and the parent of this synset.

Similar to the current version of Princeton WordNet, in RuWordNet class-instance relations are established. By now, they had been generated semi-automatically for geographical synsets for indication of their types. The part-whole relations from RuThes were semi-automatically transferred and corrected according to traditions of WordNet-like resources.

Adjectives in RuWordNet similarly to German or Polish wordnets (Gross and Miller, 1990; Maziarz et al., 2012; Kunze and Lemnitzer, 2010) are connected with hyponym-hypernym relations. Antonyms relations were transformed from association relations and currently are established between synsets, not between lexical units. Verb synsets additionally have cause and entailment relations.

Besides, to overcome so-called ”tennis problem” (Miller et al., 1990), the domain system was introduced in RuWordNet. The tennis problem is that synsets from the same domain (*tennis player, racket, court*) are very far from each other in the WordNet hierarchy. The WordNet domain system proposed in (Magnini and Cavaglia, 2000) was adapted for the RuWordNet synsets. Then domain labels were semi-



Посадить 1

Синсет

1 САДИТЬ, САЖАТЬ, САЖАТЬ В ЗЕМЛЮ, САЖАТЬ РАСТЕНИЯ, НАСАДИТЬ, **ПОСАДИТЬ** , ПОСАДИТЬ В ЗЕМЛЮ, ПОСАДИТЬ РАСТЕНИЕ, НАСАЖИВАТЬ, ПРОИЗВЕСТИ ПОСАДКУ, ПРОИЗВОДИТЬ ПОСАДКУ [Понятие РуТез: ПОСАДКА РАСТЕНИЙ]

домен

1 БОТАНИКА [Понятие РуТез: БОТАНИКА]

гипероним

1 НАЙТИ МЕСТО, ПОМЕЩАТЬ, ВОДВОРИТЬ, ВОДВОРЯТЬ, ПОМЕСТИТЬ [Понятие РуТез: ПОМЕСТИТЬ, НАЙТИ МЕСТО]

гипоним

1 ДОСАДИТЬ, ПОДСАДИТЬ, ДОСАЖИВАТЬ, ПОДСАЖИВАТЬ [Понятие РуТез: ДОСАДИТЬ, ПОДСАДИТЬ]

2 УКОРЕНИТЬ, УКОРЕНЯТЬ, УКОРЕНИТЬ РАСТЕНИЕ [Понятие РуТез: УКОРЕНИТЬ РАСТЕНИЕ]

3 ПОСАДИТЬ ЗЕЛЕНЫЕ НАСАЖДЕНИЯ, ОЗЕЛЕНИТЬ, ОЗЕЛЕНЯТЬ [Понятие РуТез: ОЗЕЛЕНИТЕЛЬНЫЕ РАБОТЫ]

4 СЕЯТЬ, СЕЯТЬ СЕМЕНА, САЖАТЬ СЕМЕНА, ВЫСЕЯТЬ, ЗАСЕЯТЬ, НАСЕЯТЬ, ПОСЕЯТЬ, ЗАСЕЯТЬ СЕМЕНА, ВЫСЕВАТЬ, ЗАСЕВАТЬ, НАСЕВАТЬ, ЗАСЕВАТЬ СЕМЕНА, ВЫСЕИВАТЬ, ЗАСЕИВАТЬ, НАСЕИВАТЬ, ЗАСЕВАТЬСЯ [Понятие РуТез: ПОСАДКА СЕМЯН]

5 САЖАТЬ НА ДРУГОЕ МЕСТО, ПОСАДИТЬ НА ДРУГОЕ МЕСТО, ПЕРЕСАДИТЬ, ПЕРЕСАДИТЬ НА ДРУГОЕ МЕСТО, ПЕРЕСАЖИВАТЬ, ПЕРЕСАЖИВАТЬ НА ДРУГОЕ МЕСТО [Понятие РуТез: ПЕРЕСАДКА РАСТЕНИЙ]

Figure 6: Screen of RuWordNet online version

automatically assigned to the RuWordNet synsets. The domains labels were represented as categories in the RuThes interface for creating knowledge-based categorization systems (see Section 5).

Fig. 6 presents description of Russian verb ”posadit” in RuWordNet. It has seven senses, the first sense corresponds to English *to plant seeds, seedlings, or plants*. The description includes synonyms and variants in the synset, the reference to the source RuThes concept *Plant setting*, the link to the domain concept (Botany), the hypernym synset (to put, locate), and hyponym synsets.

7. Conclusion

In this paper we presented the RuThes family of Russian thesauri intended for natural language processing and information retrieval applications. RuThes-like thesauri include, besides RuThes, Sociopolitical thesaurus, Security Thesaurus, and Ontology on Natural Sciences and Technology. The RuThes format is based on three approaches for developing computer resources: Princeton WordNet, information-retrieval thesauri, and formal ontologies.

The published version of RuThes thesaurus (RuThes-lite 2.0) became a basis for semi-automatic generation of RuWordNet, WordNet-like thesaurus for Russian. Currently researchers can use both RuThes-lite or RuWordNet and compare them in their applications. Other RuThes-like resources are being prepared to publication.

Acknowledgements

This work is partially supported by the Russian Foundation for Basic Research (project 16-29-09606) and the Ministry of Education and Science of the Russian Federation (project N 14.601.21.0018).

References

Azarova, I., Braslavsky, P., Zakharov, V., Kiselev, Y., Ustalov, D., and Khohlova, M. (2016). Integration of thesauri russnet and yarn. In *Proceedings of Conference ”Internet and Modern Society”*, pages 7–13.

Azarowa, I. (2008). Russnet as a computer lexicon for russian. *Proceedings of the Intelligent Information systems IIS-2008*, pages 341–350.

Balkova, V., Suhonogov, A., and Yablonsky, S. (2008). Some issues in the construction of a russian wordnet grid. In *Proceedings of the Forth International WordNet Conference, Szeged, Hungary*, volume 44.

Braslavski, P., Ustalov, D., Mukhin, M., and Kiselev, Y. (2016). Yarn: Spinning-in-progress. In *Proceedings of the Eight Global Wordnet Conference*, pages 58–65.

Clarke, D. and Zeng, M. L. (2012). From iso 2788 to iso 25964: The evolution of thesaurus standards towards interoperability and data modelling. *Information Standards Quarterly (ISQ)*, 24(1).

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., and Broda, B. (2008). Words, concepts and relations in the construction of polish wordnet. In *Proceedings of the Global WordNet Conference, Seged, Hungary*, pages 162–177.

Dobrov, B. and Loukachevitch, N. (2006). Development of linguistic ontology on natural sciences and technology. In *Proceedings of Linguistic Resources and Evaluation Conference*, pages 1077–1082.

Fellbaum, C., Ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gelfenbeyn, I., Goncharuk, A., Lehelt, V., Lipatov, A., and Shilo, V. (2003). Automatic translation of wordnet semantic network to russian language. In *International Dialog 2003 Workshop*, pages 148–154.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928.

Guarino, N. and Welty, C. A. (2009). An overview of ontoclean. In *Handbook on ontologies*, pages 201–220. Springer.

Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer.

Guarino, N. (1998). Some ontological principles for designing upper level lexical resources. In *Proceedings of First International Conference on Language Resources and Evaluation LREC-1998*, pages 28–30.

Guizzardi, G. (2011). Ontological foundations for conceptual part-whole relations: the case of collectives and their parts. In *Advanced Information Systems Engineering*, pages 138–153. Springer.

Koeva, S. (2010). Bulgarian wordnet–current state, applications and prospects. *Bulgarian-American Dialogues*, pages 120–132.

Kunze, C. and Lemnitzer, L. (2010). Lexical-semantic and conceptual relations in germanet. *Lexical-semantic relations: Theoretical and practical perspectives*, (28):163–183.

Kupriyanov, V., Kossilov, A., Maximov, N., and Kupriyanova, I. (2016). *A Semantic-Based Approach for Preserving Operational Experience of Nuclear Installations*. Technical report.

Lenat, D., Miller, G., and Yokoi, T. (1995). Cyc, wordnet, and edr: critiques and responses. *Communications of the ACM*, 38(11):45–48.

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Loukachevitch, N. and Dobrov, B. (2014). Ruthes linguistic ontology vs. russian wordnets. In *Proceedings of Global WordNet Conference GWC-2014*.

Loukachevitch, N. and Dobrov, B. (2015). The sociopolitical thesaurus as a resource for automatic document processing in russian. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 21(2):237–262.

Loukachevitch, N., Dobrov, B., and Chetviorkin, I. (2014). Ruthes-lite, a publicly available version of thesaurus of russian language ruthes. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue, Bekasovo, Russia*, pages 340–349.

Loukachevitch, N., Shevelev, A., Mozharova, V., Dobrov, B., and Pavlov, A. (2017). Ruthes thesaurus in detecting russian paraphrases. In *Conference on Artificial Intelligence and Natural Language*, pages 242–256. Springer.

Loukachevitch, N., Lashevich, G., and Dobrov, B. (2018). Comparing two thesaurus representations for russian. In *Proceedings of Global WordNet Conference GWC-2018*.

Magnini, B. and Cavaglia, G. (2000). Integrating subject field codes into wordnet. In *LREC*, pages 1413–1418.

Magnini, B. and Speranza, M. (2002). Merging global and specialized linguistic ontologies. *Proceedings of Ontolex 2002*, pages 43–48.

Mdivani, R. (2013). Thesauri of the isiss ras for social sciences and humanities. *Scientific and Technical Information Processing*, 40(3):137–141.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Nirenburg, S. and Raskin, V. (2004). *Ontological semantics*. Mit Press.

Nirenburg, S. and Wilks, Y. (2001). What's in a symbol: ontology, representation and language. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(1):9–23.

NISO. (2005). *Z39.19 - Guidelines for the Construction, Format and Management of Monolingual Thesauri*. NISO.

Pease, A. (2011). *Ontology: A practical guide*. Articulate Software Press.

Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197.

Veale, T. and Hao, Y. (2008). A context-sensitive framework for lexical ontologies. *The Knowledge Engineering Review*, 23(1):101–115.

Wilks, Y. (2009). Ontotherapy, or how to stop worrying about what there is. *Recent advances in natural language processing V*, pages 1–20.

Will, L. (2012). The iso 25964 data model for the structure of an information retrieval thesaurus. *Bulletin of the Association for Information Science and Technology*, 38(4):48–51.