CHAPTER 18

RuThes Thesaurus for Natural Language Processing

Natalia Loukachevitch and Boris Dobrov

18.1 Introduction

In natural language processing (NLP) and information retrieval (IR), it is often useful to utilize various types of knowledge, including lexical knowledge about relations between words, their senses, domain-specific knowledge, and commonsense knowledge. The conventional way to represent this knowledge within NLP systems are the so-called thesauri (= thesauruses). In NLP and IR domains, a thesaurus is a language or terminological resource describing relations between lexical or terminological units in a formalized form (in form of links), which makes it possible to use such descriptions in computer text processing.

There exist two well-known paradigms of thesauri used in computer information systems. The first paradigm is information retrieval thesauri, designated for improving document search in information retrieval systems. The role of such thesauri in information retrieval was most significant during the 1960–1980s of the twentieth century. Currently, global search engines do not use manually created thesauri. Nevertheless, the importance of such resources continues to be quite high, because such thesauri are used in information services of large international organizations as a source of recommended keywords for document indexing and search. However, these thesauri are not intended for automatic procedures of indexing and information search (ISO-25964 2011; NISO 2005).

N. Loukachevitch (✉) • B. Dobrov
Lomonosov Moscow State University, Moscow, Russia

D. Gritsenko et al. (eds.), *The Palgrave Handbook of Digital Russia Studies*, https://doi.org/10.1007/978-3-030-42855-6_18

Another paradigm of thesaurus-like resources is implemented in Princeton WordNet, created for the English language (Fellbaum 1998; Miller 1998). Since its appearance, WordNet has attracted a lot of attention of researchers and other specialists in natural language processing and information retrieval. WordNet-like thesauri (wordnets) have been initiated for many languages in the world (Vossen 1998; Bond and Foster 2013; Maziarz et al. 2016). In contrast to information retrieval thesauri, which are created for specific domains, wordnets usually represent the lexical system of a specific language in the form of sets of synonyms and relations between them.

As a detailed formalized description of the language lexical system, WordNet is used in numerous applications as a tool for automatic text processing, as a basis for generating new computational resources (e.g., ImageNet [Mishkin et al. 2017] or SentiWordNet [Baccianella et al. 2010]). But WordNet's structure is not convenient for describing the conceptual system of a broad domain because of WordNet's orientation to representing the lexical system of the language including parts of speech, lexical relations (synonyms, antonyms, derivation, etc.), and language registers (Loukachevitch and Dobrov 2014).

In this chapter, we describe the Russian thesaurus RuThes, which has been created as a tool for automatic document processing of contemporary news texts, newspaper articles, and legal texts to enable their search, categorization, clustering, and so on. In its structure, RuThes combines approaches for language and knowledge representation that are accepted in information retrieval thesauri and WordNet-like resources. The development of RuThes began more than 20 years ago. The thesaurus continues to be updated with novel concepts, words, senses, and multiword expressions, which represent the current state of the Russian language used in contemporary texts. RuThes stores knowledge about current social and political life in Russia, which can be described using the thesaurus' relations. We compare the RuThes structure with other thesaurus paradigms and provide several examples of recently introduced concepts.

The chapter is structured as follows. In Sect. 18.2 we describe the main methodologies for creating large thesauri for natural language processing and information retrieval. In Sect. 18.3, we discuss the approach to knowledge representation in the RuThes thesaurus. Section 18.4 is devoted to the description of current social and political concepts in RuThes.

18.2 Thesauri in NLP an IR

18.2.1 WordNet Thesaurus and Wordnets

The structure of Princeton University's WordNet (and other wordnets) is based on sets of synonyms—synsets. Most synsets are provided with a "gloss" explaining their meaning. If a word has several meanings, it is included into several synsets. Synset is considered by the authors as a representation of the lexicalized concept of the English language. The current WordNet (version 3.0) covers approximately 155,000 unique words and phrases, organized into

117,000 synsets. Each synset has relations with other synsets, such as hyponyms (more specific words), hyperonyms (more general words), meronyms (parts), holonyms (wholes), and others. The WordNet thesaurus includes the words of four parts of speech (nouns, adjectives, verbs, and adverbs) and is divided into four lexical nets according to these parts of speech. The synsets of each part of speech in WordNet have their own sets of relationships. Also, specific words in synsets can have their own lexical relations (antonyms, derivation). Princeton WordNet (Fellbaum 1998; Miller 1998) is freely available on the Internet (WordNet 2019), and on its basis thousands of experiments in the field of information retrieval and natural language processing were carried out (for more on linguistic resources, see Chaps. 29, 19 and 26).

Bond et al. (2016) noted that WordNet-like resources (wordnets) created for different languages, while preserving the basic structure of WordNet, can differ significantly from each other in terms of the inclusion of words and expressions in synsets, the use of semantic relations between synsets, and the interpretation of specific semantic relations. Also, in wordnets, approaches to the description of polysemy can vary considerably, which leads to a more fine or coarse system of representing the senses of ambiguous words. There may be different approaches to the inclusion of multiword expressions into wordnets.

Some features of the WordNet structure are not very convenient for describing the conceptual system of a specific domain. These features include sets of synonyms (synsets) as a basic unit of the thesaurus, the division into part-of-speech structures, lexical relations, and approaches to inclusion phrases. However, several attempts to create domain-specific wordnets (e.g., ArchiWordNet, Jur-WordNet) have been made (for a review, see Lüngen et al. 2008).

18.2.2 Information Retrieval Thesauri

Information retrieval thesauri are important instruments in information and library services; for years, they were used for representing the domain knowledge in information retrieval systems. International and national standards have been published in the 1980s and continue to be updated (ISO-25964 2011; NISO 2005; Dextre Clarke and Zeng 2012). There exist some very influential international thesauri such as EUROVOC—the thesaurus of the European Community (EUROVOC Thesaurus 1995), the UNBIS thesaurus of the United Nations (United Nations 1976), the Art and Architecture thesaurus (Art & Architecture Thesaurus Online 2018) and others.

Information retrieval thesauri are less known and utilized for NLP purposes because they are intended to be used only in manual or automated indexing by human indexers, according to the thesaurus standards (ISO-25964 2011; NISO 2005). However, the principles of describing broad and complex domains are important for comparison with the WordNet structure.

The main units of information retrieval thesauri are domain terms denoting domain concepts. Domain concepts can have several variants of text representation, which are considered as synonyms. Among synonyms, the most representative variant, called a descriptor or preferred term, is chosen. Other terms included in the thesaurus are called nonpreferred terms and used as auxiliary units helping to find preferred terms.

Every descriptor should be formulated unambiguously. If a clear and unambiguous descriptor cannot be formulated, the term that is taken as a descriptor is supplied with a relator (a short label) or comment. In standards, there are special guidelines for introducing multiword descriptors (NISO 2005). The set of the thesaurus descriptors should be sufficient to describe the topics of the absolute majority of the documents in the domain. To explain why such thesauri are not suited for use in automatic document processing, we would like to provide several examples from the EUROVOC thesaurus (EUROVOC 1995). EUROVOC is created for 23 languages of the European Union and therefore it does not include Russian, but this thesaurus is one of the most well-known resources and therefore its principles are important to consider.

To improve the domain representation for humans, the guidelines for the creation of information retrieval thesauri often recommend not to include certain kinds of terms in a thesaurus (infrequent terms, terms that are too specific, similar terms etc.; United Nations 2009). Relying on human indexers, traditional information retrieval thesauri try to limit the inclusion of ambiguous terms, which leads to problems in automatic document processing. In EUROVOC, for example, the single-word term *bank* is presented in only one sense; other senses are described in form of multiword terms (*sperm bank, data bank, blood bank*). Note, that in WordNet, the word *bank* has ten senses as a noun and eight senses as a verb. In the defense category, EUROVOC does not contain such terms as *soldier* or *military force*; only the descriptor *armed force* is presented.

The relations in information retrieval thesauri are quite different from WordNet-like lexical relations. Information retrieval thesauri have a small set of generalized relations, which are usually subdivided into two classes: hierarchical and associative. The most frequent type of hierarchical relations between preferred terms in information retrieval thesauri are the broader-narrow relations (BT and NT relations), comprising class-subclass, instance-class, and sometimes part-whole relationships. The associative relations convey various other types of domain-specific relations between concepts (related term (RT) relation). The standards and manuals on thesaurus development formulate principles for representing associative relations as the most significant ones (NISO 2005; Aitchinson and Gilchrist 1987).

The RT relations are considered to be symmetric, but looking at the existing thesauri, it is possible to see that this is not true in many cases. For example, in EUROVOC the *air transport* descriptor has RT relations with such descriptors as *air law*, *air traffic control*, and *aviation fuel*, which are much narrower than the *air transport* descriptor. This simple system of relations has been criticized

in many works (Tudhope et al. 2001) but it has an important advantage: it can be applied to any domain without additional efforts to develop the detailed set of domain-specific relations, which always is a very complex task.

In Russia, the most known information retrieval thesauri are developed in the Institute of Scientific Information of Russian Academy of Sciences (INION RAN). This institution publishes separate issues of thesauri on economics, sociology, linguistics, and so on, created according to the guidelines of international and national standards on thesaurus construction. These thesauri also cannot be used for automatic processing of document and news flows (Mdivani 2013).

18.3 RuThes Structure, Units, and Relations

18.3.1 RuThes General Structure

In the construction of RuThes, both popular paradigms for computer thesauri were used: concept-based units, a small set of relation types, and rules for including multiword expressions as in information retrieval thesauri; language-motivated units, detailed sets of synonyms, and description of ambiguous words as in wordnets. Also, some issues of ontology research—for example, concepts as main units, strictness of relation description, necessity for many-step inference—are accounted for (Guarino 1998, 2009).

RuThes is a hierarchical network of concepts. Each concept has a name, relations with other concepts, and a set of language expressions (words, phrases, terms) whose meanings correspond to the concept. The whole set of RuThes' concepts is subdivided into general lexicon and sociopolitical thesaurus. *General Lexicon* comprises general concepts and words that can be met in various specific domains such as *sozdanie* (creation), *udalit'* (remove), *uslovnye* (conditional). *Sociopolitical Thesaurus* contains thematically oriented lexemes and multiword expressions as well as domain-specific terms of the broad sociopolitical domain. The whole RuThes thesaurus includes more than 60,000 concepts and more than 200,000 Russian text entries (words and expressions). The published version of RuThes for use in noncommercial applications includes 110,000 text entries (RuThes 2019).

The *sociopolitical domain* is the domain of problems, relationships, and situations of the contemporary society (Loukachevitch and Dobrov 2015). Subdomains of the sociopolitical domain are themselves large domains such as economics, law, or international relations, each with its own terminology. However, the specific feature of the sociopolitical domain (and its subdomains) is that most domain terms are known to nonprofessionals. Here, in the sociopolitical domain, the general language and domain terminologies adjoin and mix with each other. At present, the RuThes sociopolitical thesaurus includes terminology from such domains as politics, elections, sociology, demography, social security, civil and criminal law, the court system, banking, security, economics (including macroeconomics, industry, agriculture, and transport), ecology, accidents, sports, culture, and others.

18.3.2 RuThes Units

The RuThes thesaurus is a hierarchy of concepts viewed as units of thought. A concept is associated with the set of language expressions that refer to it in texts. This approach is similar to approaches of traditional information retrieval construction (NISO 2005). In most cases, concepts should have denotational distinctions from related concepts. Such distinctions can be expressed in a specific set of relationships or associated language expressions: *text entries*.

Words and phrases whose meanings refer to the same concepts represented in the thesaurus are called ontological synonyms. Ontological synonyms can comprise sense-related words belonging to different parts of speech (i.e., *privatizaciâ* [privatization] *vs. privatizirovat'* [to privatize]); in contrast to traditional terminological resources and information retrieval thesauri that contain mainly nouns or noun phrases. A thesaurus for automatic document processing should contain various types of language units. Also, language expressions relating to different linguistic styles, technical terms, and lexical units can be presented as ontological synonyms related to the same concept. For example, the concept *Oil industry* has the following text entries: *neftânaâ promyšlennost'* (oil industry)—neutral, *neftânka*—slang, *nefteprom*—abbreviation. Compositional multiword expressions may be included into synonymic sets as well. Each concept should have a clear, univocal, and concise name. Such names often help to express and delimit the denotational scope of the concept. In addition, the concepts' names can be used in the analysis of the results of automatic document analysis, for example in visualization of trends or as cluster names.

Ontological synonyms, variants of lexical units, and technical terms (Nazarenko and Zargayouna 2009) are collected specially. After a concept has been introduced, an expert searches for all possible synonyms or orthographic variants, single words, and phrases that can be associated with it. These synonymic sets can also include multiple variants of the references to the same concept. For example, the concept *Ohrana prirody* (Nature protection) is associated with almost 50 different text entries in Russian, for example *zaŝita prirody* (defense of nature), *sohranenie prirody* (maintenance of nature), *zaŝiŝat' prirodu* (to protect nature), *sohranât' prirodu* (to maintain nature), and others. These variants are useful to describe in the thesaurus because they directly refer to their concept. Besides, multiword term variants often contain ambiguous words within themselves. Thus, the inclusion of such term variants decreases the overall lexical ambiguity and facilitates disambiguation. All variants are collected during the analysis of real texts, usually news articles, legislative acts, or domain-specific documents.

In fact, the introduction of such a concept as *Nature protection* corresponds more to information retrieval thesauri than wordnets, because one of the important principles of WordNet-like resources is to include single words and lexicalized phrases into synsets (Bentivogli and Pianta 2004; Maziarz and Piasecki 2018). The phrase *nature protection* seems compositional, but the

concept *Nature protection* is significant for the contemporary life of the society and it has relations with other important concepts of the sociopolitical domain.

As can be seen, one of the difficult issues in developing application-oriented resources, such as wordnets or information retrieval thesauri is the inclusion of units (synsets or descriptors) based on the senses of multiword expressions, for example noun compounds (Bentivogli and Pianta 2004). Manuals and standards for information retrieval thesaurus development provide detailed principles for multiword term selection (NISO 2005; Aitchinson and Gilchrist 1987). In RuThes, the introduction of concepts based on multiword expressions is not restricted but encouraged if this concept adds some new information to the knowledge described in the thesaurus (Loukachevitch and Lashevich 2016).

18.3.3 RuThes Relations

Conceptual relations in the thesaurus may be utilized for several purposes, including query expansion in information retrieval, clustering related concepts mentioned in a text as a basis for better recognition of the main theme and subthemes in the document, and disambiguation of ambiguous terms and lexical units. Working with such a broad scope of concepts, we utilize a set of relations that can be applied to concepts in various domains, in contrast to domain-dependent relations.

RuThes has a small set of conceptual relations consisting of four main relations that describe the most important links of a concept. In fact, the current set of relations in the thesaurus is a more ontologically motivated variant of classic inter-descriptor relationships in information retrieval thesauri, which usually include hierarchical relations, such as broader term (BT) and narrower term (NT), and associative relations—related term (RT).

The first relation of RuThes is *the class-subclass relation* as it is treated in ontological approaches (Guarino 1998; Gangemi et al. 2003). To establish such relations, we apply tests similar to those used in ontology development. The tests are directed toward avoiding incorrect use of class-subclass relations and not mixing them up with other types of relations (such as type-role relation, class-instance relation), because errors in relation types degrade logical inference (Gangemi et al. 2003). The class-subclass relationship is considered as a transitive relation with the inheritance property.

The second relationship is *part-whole relation*, which is established using specific ontological restrictions (Gangemi et al. 2003). Our decision on part-whole relations is based on the following principles:

- Broad treatment of part-whole relations from the semantic point of view,
- Restriction of ontological subtypes of part-whole relations,
- Postulating the transitivity of part-whole relations.

Part-whole relations in RuThes comprise such relationships as parts of physical objects, territorial and geographical parts, process parts, and others (see examples in Table 18.1). Also, some other relationships are presented as part-whole relations in RuThes: an attribute and its bearer, a role or a participant in the situation (Winston et al. 1987, 27–28), entities and situations in the encompassing sphere of activity (Table 18.1).

In such a broad scope, part-whole relations described in RuThes are close to the so-called *internal relations* (parthood, constitution, quality inherence, and participation) as described by Guarino (2009). At the same time, part-whole relations in RuThes have a very important restriction (correlating with the information retrieval thesauri guidelines about the necessity to describe only inherent properties as hierarchical relations [NISO 2005]): a concept-part should be related to its whole during the normal existence of its instances: the so-called *ontological dependence*.

To analyze the ontological dependence between entities X and Y, it is necessary to determine whether entity X can exist by itself or whether its existence depends on the existence of Y. We describe the following types of dependent parts in RuThes:

Table 18.1 Types and examples of part-whole relations in RuThes

Type of relationship	*Part*	*Whole*
Parts of physical objects	*starter dvigatelá* (motor starter) *kost'* (bone)	*dvigatel' vnutrennego sgoraniá* (internal combustion engine), *skelet* (skeleton)
Territorial and geographical parts	*oazis* (oasis) *izbiratel'nyj učastok* (electoral precinct) *bankovskij sejf* (bank safe)—	*pustynâ* (desert), *izbiratel'nyj okrug* (electoral district), *bankovskoe hraniliše* (bank vault)
Process parts	*izbiratel'naâ tehnologiâ* (electoral technology)	*predvybornaâ kampaniâ* (pre-election campaign)
Text and musical parts	*vvedenie* (text introduction) *muzykal'nyj interval* (musical interval)	*tekst* (text), *muzykal'naâ kompoziciâ* (musical composition)
Members	*člen političeskoj partii* (political party member) *deputat Gosudarstvennoj Dumy* (State Duma Deputy)	*političeskaâ partiâ* (political party), *Gosudarstvennaâ Duma* (State Duma, the lower house of the Russian Parliament)
Substance as a part	*židkost' v organizme* (body fluids)	*telo* (body of living organism)
An attribute and its bearer	*skorost'* (speed) *glasnost' vyborov* (election publicity)	*dviženie* (movement), *vybory* (election)
Roles and participants in a situation	*investor* (investor) *igrok* (player)	*investirovanie* (investing), *igra* (game)
Entities and situations in the encompassing sphere of activity	*zavod* (industrial plant) *sportsmen* (sportsman)	*promyšlennost'* (industry), *sport* (sport)

- Inseparable part, which is a part that cannot exist without its whole, such as *oazis* (oasis)—*pustynâ* (desert);
- Mandatory whole, when a part requires the existence of at least one entity described as a whole, such as *bankovskij sejf* (bank safe) and *bankovskoe hraniliŝe* (bank vault) (Guizzardi 2011).

Thus, we put existential constraints on the part-whole relations in RuThes. These constraints do not change the transitivity of part-whole relations if it was postulated. The inference mechanism can thereby utilize the transitivity of part-whole relations and rely on the chain of part-whole relations (Guizzardi 2011; Loukachevitch and Dobrov 2015).

The final types of relationships are *nonsymmetrical and symmetrical associations*, which are subdivided from the symmetric related term (RT) relation of conventional information retrieval thesauri. The nonsymmetrical associations are established on the basis of the ontological dependence of concepts. Symmetrical associations are described in the very restricted number of cases.

Associative relationships (RT relations) are quite common in information retrieval thesauri; they are established to provide additional links between descriptors for use in the indexing or retrieval of documents (NISO 2005). Such relations in information retrieval thesauri are always considered as symmetrical; however, many associative relations found in published thesauri demonstrate the evident absence of symmetry, for example *illness—disease prevention*, *illness—sick leave* (EUROVOC), et cetera. The first term in each pair is much more general than the other one.

Considering the problems involved in formalizing traditional information retrieval thesauri to adapt them to the contemporary level of ontological research, some authors propose changing the thesaurus's traditional system of relations to a formalized set of predicates and to provide axioms for such a set (Soergel et al. 2004). However, in creating such multidomain resources as RuThes, it is very difficult to find the universal set of semantic relations and apply them consistently. Therefore, we substituted the traditional thesaurus relation of symmetric association with another quite generalized relation, which can be applied in many various domains. We usually refer to this relation as a nonsymmetrical association, asc_1–asc_2. The definition of this relation is again based on a variant of ontological dependence, the so-called *external dependence* in ontological terms (Gangemi et al. 2003; Guarino 2009). This relation is established between two concepts c_1 and c_2 when two requirements are fulfilled:

- Neither class-subclass nor part-whole relations can be established between c_1 and c_2 in the thesaurus.
- The following assertion is true: "concept c_2 exists" means "concept c_1 exists" (necessarily existent entities are excluded from consideration).

Table 18.2 Examples of conceptual dependence relations denoted as nonsymmetrical associations in RuThes

Type of relationships	*Main concept*	*Dependent concept*
Instrument—professional that uses this instrument	*skripka* (violin)	*skripač* (violinist)
Entity—branch of science that studies such entities	*životnoe* (animal)	*zoologiá* (zoology)
	serdce (heart)	*kardiologiá* (cardiology)
Entity and related entity	*bagaž* (luggage)	*bagažnaá karusel'* (luggage carousel)
Entity and actions that applied to these entities	*krov'* (blood)	*donorstvo krovi* (blood donation)
	eda (food)	žarka (frying)
Entity and its specific problems	*les* (forest)	*lesnoj požar* (forest fire)
	serdce (heart)	*bolezn' serdca* (heart disease)
Entity and opposing entity or action	*virus* (virus)	*antivirus* (antivirus)

These two conditions mean that the concept c_2 (dependent concept) externally depends on $c_1 : asc_1(c_2, c_1) = asc_2(c_1, c_2)$. Table 18.2 presents some examples of conceptual relationships, where conceptual dependence can be seen.

Relations of ontological dependence are applicable to various domains; therefore, they are usually used in top-level ontologies (Gangemi et al. 2003). An additional advantage of using these relations in thesauri for automatic document processing is their usefulness for describing links between a concept based on the sense of a compositional multiword expression and concepts corresponding to the components of this multiword expression. As a result, a multiword-based concept (e.g., *Automobile racing*) is described as the dependent concept and its component concept (*Automobile*) as the main concept. This allows us to introduce concepts based on various types of multiword expressions and to establish their necessary relations.

18.4 Description of Social and Political Concepts in RuThes

The specific part of RuThes called Sociopolitical thesaurus provides detailed coverage of thematic lexical units and terms in the broad sociopolitical domain of contemporary written Russian (mainly news articles, laws, and official documents). The thesaurus was utilized in document-processing applications within information,retrieval and information analytical systems (Loukachevitch and Dobrov 2015). Every project gave the opportunity to improve the descriptions of lexical senses, reveal useful expressions, and add domain terms of new subdomains of the sociopolitical domain, which, in turn, improved the description of related lexical senses.

Let us consider several examples of recently introduced concepts related to popular topics discussed in the Russian and international press and their descriptions in RuThes. Figure 18.1 represents the description of concepts

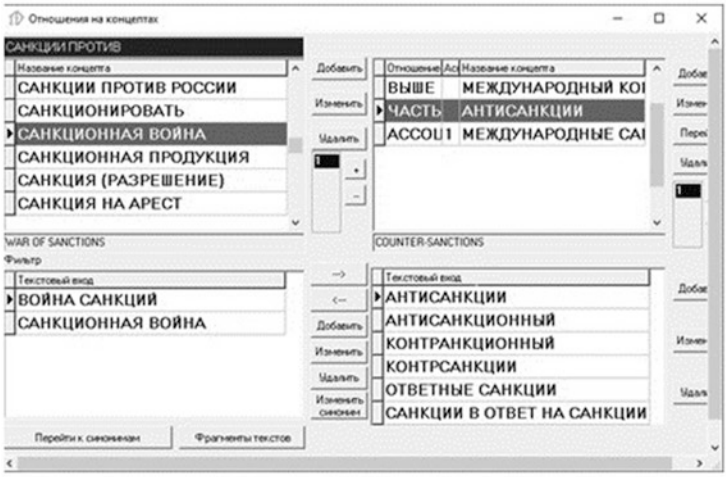

Fig. 18.1 Representation of the current international sanctions situation in the thesaurus form

related to sanctions: *Sankcii protiv Rossii* (Sanctions against Russia), *Sankcionnaâ vojna* (War of sanctions), *Sankcionnaâ produkciâ* (Products under sanctions). The upper-left form enumerates a list of concepts in alphabetical order.

The left-lower form shows Russian text entries for the concept *War of sanctions* such as *vojna sankcij* (war of sanctions) and *sankcionnaâ vojna* (sanctions war). The right-upper form presents the relations of the highlighted concept. Figure 18.1 shows the relation of the *War of sanctions* concept with such concepts as *Meždunarodnyj konflikt* (International conflict), *Antisankcii* (Counter-sanctions), and *Meždunarodnye sankcii* (International sanctions). In particular, the *War of sanctions* concept is described as dependent from the concept *International sanctions*, because it could not appear without this concept. The *Counter-sanctions* concept is described as a part of *War of sanctions*. The lower-right form shows Russian text entries for the related concept *Antisankcii* (Counter-sanctions). They include: nouns (*antisankcii, kontrsankcii*), noun groups (*otvetnye sankcii* [sanctions as an answer]), and adjectives (*antisankcionnyj, kontrsankcionnyj*).

After the pension reform in Russia was announced in 2018, new concepts *Predpensioner* (Person before retirement age) and *Predpensionnyj vozrast* (Before retirement age) were introduced in the thesaurus. These concepts appeared in Russian law to provide social security to some categories of the population in relation to raising the retirement age. The *Before retirement age* concept is described as a part (property) of *Person before retirement age* according the thesaurus guidelines. The concept *Person before retirement age* depends on the concepts *Pensioner* and *Pension system* because it requires their existence.

An innovation of the Russian transport law introduced yellow boxes on roads, a specific kind of road marks (box marking). In Russian, the concept is called *Vafel'naâ razmetka* (literally, *waffle marking*). The concept's set of Russian text entries includes word *vafel'nica*, which previously meant only "waffle iron," a kitchen appliance for baking waffles. Therefore, the new sense of the word *vafel'nica* and new multiword expression *vafel'naâ razmetka* have been added into RuThes.

In recent years, cryptocurrencies were actively discussed. The corresponding concepts: *kriptovalûta* (cryptocurrency), *èlektronnye den'gi* (electronic money), *Bitcoin*, *kriptomat* (Cryptocurrency ATM machine) have been introduced into the thesaurus.

Thus, RuThes provides detailed coverage of thematic lexical units and terms in the broad sociopolitical domain of contemporary written Russian (mainly news articles, laws, and official documents). The thesaurus can be used as a conceptual indexing tool in information analytical systems. RuThes can also be a useful instrument for developing knowledge-based categorization systems in conditions when a training collection for machine learning methods is absent and cannot be easily created. It is possible because the thesaurus contains thousands of words and expressions stored in a hierarchical structure, which can be used in the description of categories for automatic text categorization (Loukachevitch and Dobrov 2015).

18.5 RuThes as a Source for Russian WordNet

Despite the fact that RuThes is currently published for noncommercial use, people would like to have a large Russian wordnet. Therefore, a transforming procedure from the published version of RuThes (RuThes-lite) to the largest Russian WordNet (RuWordNet 2019) has been initiated. One of the most distinctive features of WordNet-like resources is their division into synset nets according to parts of speech. Therefore, all text entries of RuThes-lite were subdivided into three parts of speech: nouns (single nouns, noun groups, and preposition groups), verbs (single verbs and verb groups), adjectives (single adjectives and adjective groups). We have obtained 29,297 noun synsets, 12,865 adjective synsets, and 7636 verb synsets. The divided synsets were linked to each other with the relation of part-of-speech synonymy.

The hyponym-hypernym lexical relations (hyponymy shows the relationship between a generic term [hypernym] and a specific instance of it [hyponym]) were established between synsets of the same part of speech. These relations include direct hyponym-hypernym relations from RuThes-lite. 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.

Other RuThes relations were modified. The part-whole relations from RuThes were semi-automatically transferred and corrected according to traditions of WordNet-like without the expanded set of part-whole relations. Some part-whole relations were transformed to domain relations, for example *zavod* synset (industrial plant) is related to the domain *promyšlennost'* (industry) via the domain relation. The ontological dependence relations of RuThes were manually transformed to appropriate semantic relations such as antonyms, cause, entailment, and some others. RuWordNet is publicly available (RuWordNet 2019).

18.6 Conclusion

In this chapter, we described the RuThes thesaurus that was created as a linguistic and terminological resource for automatic document processing in Russian. In the construction of RuThes, both popular paradigms for computer thesauri were used: concept-based units, a small set of relation types, and rules for including multiword expression as in information retrieval thesauri; language-motivated units, detailed sets of synonyms, and description of ambiguous words as in wordnets. A large part of RuThes is devoted to the description of terms and concepts related to the current sociopolitical life in Russia and in the world—the so-called Sociopolitical thesaurus.

We have supported the development of RuThes for many years by introducing new concepts, representing new senses, and recording multiword expressions. In this chapter, we have showed some examples of representing newly appeared concepts related to important internal and international events. We demonstrated how we used the thesaurus' relation system for describing these concepts. Hence, we consider RuThes as a kind of formalized encyclopedia of social and political life of the contemporary society.

References

Aitchinson, Jean, and Alan Gilchrist. 1987. *Thesaurus Construction: A Practical Manual*. London: Aslib.

Art & Architecture Thesaurus Online. 2018. The J. Paul Getty Trust. http://www.getty.edu/research/tools/vocabularies/aat/.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of Language Resources and Evaluation Conference LREC-2010* 10: 2200–2204.

Bentivogli, Luisa, and Emanuele Pianta. 2004. Extending Wordnet with Syntagmatic Information. In *Proceedings of Second Global WordNet Conference,* 47–53. Brno, Czech Republic.

Bond, Francis, and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* 1: 1352–1362.

Bond, Francis, Piek Vossen, Johm McCrae, and Christiane Fellbaum. 2016. CILI: The Collaborative Interlingual Index. *Proceedings of the 8th Global WordNet Conference 2016 (GWC2016)*, 27–30.

Dextre Clarke, Stella, and Marcia Lei Zeng. 2012. From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards Towards Interoperability and Data Modelling. *Information Standards Quarterly (ISQ)* 24 (1): 20–26

EUROVOC, Thesaurus. 1995. Vol. 1–3/European Communities. Luxembourg: Office for Official Publications of the European Communities.

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Gangemi, Aldo, Roberto Navigli, and Paula Velardi. 2003. The OntoWordNet Project: Extension and Axiomatisation of Conceptual Relations in Wordnet. In *Proceedings of International Conference on Ontologies, Databases and Applications of Semantics* (ODBASE), 820–838.

Guarino, Nicola. 1998. Some Ontological Principles for Designing Upper Level Lexical Resources. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 527–534. Granada, Spain.

———. 2009. The Ontological Level: Revisiting 30 Years of Knowledge Representation. In *Conceptual Modeling: Foundations and Applications: Essays in Honor of John Mylopoulos*, Lecture Notes in Computer Science, 5600, 52–67. Berlin and Heidelberg: Springer-Verlag.

Guizzardi, Giancarlo. 2011. Ontological Foundations for Conceptual Part-Wholes Relation: The Case of Collectives and Their Parts. In *Advanced Information Systems Engineering: Proceedings of the 23rd International CAiSE*, (London, UK, June 20–24) Lecture Notes in Computer Science, 6741, 138–153. Berlin and Heidelberg: Springer-Verlag.

INION RAN. Thesauri and Subject Headings. http://inion.ru/resources/bazy-dannykh-inion-ran/.

International Standards Organization (ISO). 2011. *ISO 25964 Information and Documentation—Thesauri and Interoperability with Other Vocabularies*.

Loukachevitch, Natalia, and Boris Dobrov. 2014. RuThes Linguistic Ontology Vs. Russian Wordnets. In *Proceedings of the Seventh Global WordNet Conference (GWC 2014)*, 154–162. Tartu, Estonia.

———. 2015. The Sociopolitical Thesaurus as a Resource for Automatic Document Processing in Russian. *Terminology. Special Issue Terminology across Languages and Domains* 21 (2): 238–263.

Loukachevitch, Natalia, and German Lashevich. 2016. Multiword Expressions in Russian Thesauri RuThes and RuWordNet. In *Artificial Intelligence and Natural Language Conference (AINL)*, IEEE, 1–6.

Lüngen, Harald, Claudia Kunze, Angelika Storrer, and Lothar Lemnitzer. 2008. Towards an Integrated OWL Model for Domain-Specific and General Language Wordnets. In *Proceedings of the 4th Global Wordnet Conference (GWC-2008)*, 281–296.

Maziarz, Marek, and Maciej Piasecki. 2018. Towards Mapping Thesauri onto plWordNet. *Proceedings of Global Wordnet Conference GWC-2018*, 45–53.

Maziarz, Marek, Maciej Piasecki, Eva Rudnicka, Stan Szpakowicz, and Pawel Kędzia. 2016. Plwordnet 3.0–a Comprehensive Lexical-Semantic Resource. In *Proceedings*

of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2259–2268.

Mdivani, Robert. 2013. Creating a Multilingual Thesaurus for the Social Sciences. Linguistic and Intercultural Problems. *Scientific and Technical Information Processing* 40 (3): 137–141.

Miller, George. 1998. Nouns in WordNet. In *WordNet—An Electronic Lexical Database*, ed. Christiane Fellbaum, 23–47. Cambridge, MA: The MIT Press.

Mishkin, Dmytro, Nikolay Sergievskiy, and Jiri Matas. 2017. Systematic Evaluation of Convolution Neural Network Advances on the Imagenet. *Computer Vision and Image Understanding* 161: 11–19.

Nazarenko, Adeline, and Haifa Zargayouna. 2009. Evaluating Term Extraction. In *Proceedings of International Conference Recent Advances in Natural Language Processing (RANLP'09)*, 299–304. Hissar, Bulgaria.

NISO. 2005. *Guidelines for the Construction, Format and Management of Monolingual Thesauri*, ANSI/NISO Z39.19. Bethesda, MD: NISO Press.

RuThes. 2019. Thesaurus for Natural Language Processing in Russian. http://www.labinform.ru/pub/ruthes/index_eng.htm.

RuWordNet. 2019. WordNet-like Thesaurus for Russian. http://ruwordnet.ru/en.

Soergel, Dagobert, Boris Lauser, Anita Liang, Frehivot Fisseha, Johannes Keizer, and Stephen Katz. 2004. Reengineering Thesauri for New Applications: The AGROVOC Example. *Journal of Digital Information Article* 4 (4). Accessed January 3, 2019. https://journals.tdl.org/jodi/index.php/jodi/article/view/112/111.

Tudhope, Douglas, Harith Alani, and Christopher Jones. 2001. Augmenting Thesaurus Relationships: Possibilities for Retrieval. *Journal of Digital Information* 1 (8). Available at https://eprints.soton.ac.uk/254484/1/Tudhope_JoDI.pdf (Accessed 4 July 2020).

United Nations. 1976. *UNBIS Thesaurus*, English edition. New York: Dag Hammarskjold Library of the United Nations.

———. 2009. Guidelines for Analysis of UN Documents and Publications. UNBISnet–United Nations Dag Hammarskjöld Library. http://www.un.org/Depts/dhl/unbisref_manual/indexpolicy/guidelines.htm.

Vossen, Piek. 1998. Introduction to EuroWordNet. In *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, 1–17. Netherlands: Springer.

Winston, Morton, Roger Chaffin, and Douglas Herrmann. 1987. A Taxonomy of Part-Whole Relations. *Cognitive Science* 11 (4): 417–444.

WordNet. 2019. https://wordnet.princeton.edu/.