# Comparing Two Thesaurus Representations for Russian

**Natalia Loukachevitch**
Lomonosov Moscow State University, Moscow, Russia
Tatarstan Academy of Sciences, Kazan, Russia
louk_nat@mail.ru

**German Lashevich**
Kazan Federal University
Kazan, Russia
design.ber@gmail.com

**Boris Dobrov**
Lomonosov Moscow State University, Moscow, Russia
dobrov_bv@mail.ru

## Abstract

In the paper we presented a new Russian wordnet, RuWordNet, which was semi-automatically obtained by transformation of the existing Russian thesaurus RuThes. At the first step, the basic structure of wordnets was reproduced: synsets' hierarchy for each part of speech and the basic set of relations between synsets (hyponym-hypernym, part-whole, antonyms). At the second stage, we added causation, entailment and domain relations between synsets. Also derivation relations were established for single words and the component structure for phrases included in RuWordNet. The described procedure of transformation highlights the specific features of each type of thesaurus representations.

## 1 Introduction

WordNet thesaurus is one of the popular language resources for natural language processing (Fellbaum, 1998). The projects for creating WordNet-like resources have been initiated for many languages in the world (Vossen, 1998; Bond and Paik, 2012). Other thesaurus models are rarely discussed, created and used in NLP.

In several works, S.Szpakowicz and co-authors (Jarmasz and Szpakowicz, 2004; Aman and Szpakowicz, 2008; Kennedy and Szpakowicz, 2008) evaluated two versions of Roget's thesaurus in several applications. Borin and colleagues (Borin and Forsberg, 2009; Borin et al. 2013) compared the structure of the Swedish thesaurus Saldo with the WordNet structure. In (Borin et al., 2014) automatic generation of Swedish Roget's thesaurus and its comparing with the existing Roget-style thesaurus for Swedish is discussed.

For the Russian language, RuThes thesaurus has been created more than fifteen years ago (Loukachevitch and Dobrov, 2002). It was 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 (Dobrov and Pavlov, 2010), text categorization (Loukachevitch and Dobrov, 2015), detecting Russian paraphrases (Loukachevitch et al., 2017), etc.

Using the RuThes model for the concept representation, several domain-specific thesauri have been created for NLP and domain-specific information-retrieval applications including Sociopolitical thesaurus (Loukachevitch and Dobrov, 2015), Ontology on Natural Sciences and Technology (Dobrov and Loukachevitch, 2006), Banking thesaurus (Nokel and Loukachevitch, 2016) and others. Currently, RuThes concepts provide a basis for creating the Tatar Socio-Political Thesaurus (Galieva et al., 2017).

In 2013, RuThes was partially published for non-commercial use (Loukachevitch et al., 2014). But people would like to have a large Russian wordnet. Therefore, we have initiated a transforming procedure from the published version of RuThes (RuThes-lite) to the largest Russian WordNet (RuWordNet[1]), which we describe in this paper. This transformation allows us to show similarities and differences between two resources in a detailed way. RuWordNet currently includes 115 thousand unique words and phrases.

[1] http://ruwordnet.ru/en/

The structure of this paper is as follows. In Section 2, we describe related work. Section 3 presents the structure of RuThes thesaurus, including the set of relations and principles of work with multiword expressions. Section 4 describes the main stages for creating the basic structure of RuWordNet. Section 5 is devoted to enrichment of the basic RuWordNet relations.

## 2 Related work

Creating large lexical resources like WordNet from scratch is a complex task, which requires effort for many years (Azarowa, 2008). To speed up the development of a wordnet for own language, the first version of such a resource can be created by automatically translating Princeton WordNet into the target language (Vossen, 1998; Gelfenbein et al., 2003; Sukhonogov et al. 2005), but then considerable effort is required to proofread and correct the obtained translation.

As an intermediate approach, researchers propose a two-stage creation of a wordnet for a new language: first translating and transferring the relations of the top concepts of Princeton WordNet (the so-called core WordNet), and then manually replenishing hierarchies based on dictionaries and text corpora. This approach was used in the creation of such resources as DanNet (Pedersen, 2010) and EuroWordNet (Vossen, 1998).

After analyzing the existing approaches to the development of wordnets, the creators of the Finnish wordnet (FiWN) decided to translate Princeton WordNet manually, using the work of professional translators. As a result, the Finnish wordnet was created on the basis of translation of more than 200 thousand word senses of Princeton WordNet words within 100 days (Lindén and Niemi, 2014).

In work (Braslavsky et al., 2012), it was proposed to develop a new Russian wordnet (YARN) using the Russian Wiktionary and crowdsourcing. The authors planned to attract a large number of students and interested people to create a new resource.

There are at least four known projects for creating a wordnet for the Russian language. In RussNet (Azarova et al., 2004), the authors planned to create the Russian wordnet from scratch, guided by the principles of Princeton WordNet. In two different projects described in (Gelfenbein et al., 2003; Sukhonogov et al. 2005), attempts were made to automatically translate WordNet into Russian, with all the original thesaurus structure preserved. The results of (Gelfenbein et al., 2003) are published, but the analysis of the thesaurus generated in this way shows that it requires considerable editing or the use of better algorithms.

The last project YARN (Yet Another Russian wordNet) was initiated in 2012 and initially was created on the basis of crowdsourcing, i.e. participation in the work of filling the thesaurus by a large number of participants. Currently, YARN contains a significant number of synsets with a small number of relationships between them. The published version[2] of the YARN thesaurus contains too many similar or partially similar synsets.

In (Azarova et al., 2016), the authors describe the project on the integration of the thesaurus RussNet (Azarowa., 2008) and the thesaurus YARN (Braslavsky et al., 2012) into a single linguistic resource, where the expert approach and the crowdsourcing will be combined.

In (Khodak et al., 2017), a new approach to automatic wordnet construction is presented and tested on a specially prepared Russian dataset comprising senses of 600 words (200 nouns, 200 verbs, and 200 adjectives). The approach is based on translation of English synsets, and a number of techniques of clustering and assessing the obtained translation. For Russian, the authors report 60% F-measure on the above-mentioned tests. However, the analysis of the dataset showed that the presented Russian words have much more senses than it is usually presented in Russian dictionaries. For example, word *опасность (danger)* is usually described as having 2 senses. But in the dataset it has 6 senses. Word *оборудование (equipment)* is usually described with 2 senses, but in the dataset it has 8 senses. It looks that the expert labeling of Russian senses for the dataset was somehow biased to English and its representation in Princeton WordNet.

## 3 RuThes Structure and Relations

RuThes (Loukachevitch and Dobrov, 2014; Loukachevitch et al., 2014) and WordNet are both thesauri, i.e. lexical resources in that words similar in meaning are gathered into synsets (WordNet) or concepts (RuThes), between which relations are established. When applying the two thesauri to text processing, similar steps should be carried out, including a comparison of the text

[2] https://russianword.net/

with the thesaurus, and the use of the described relations if necessary. There are also significant differences between the thesauri.

Firstly, in RuThes there is no division into lexical networks by parts of speech. Any part of speech can be associated with the same RuThes concept, if they mean the same (so-called part-of-speech synonyms). Each thesaurus concept has a unique name.

To provide morpho-syntactic information for a word, each RuThes entry has parts of speech labels. The morpho-syntactic representation of a multiword expression contains the syntactical type of the whole group, the head word, parts of speech and lemmatized forms for each component word.

Therefore, secondly, when establishing relations in RuThes, it is often impossible to apply synonym tests based on the interchangeability of words in different contexts (Miller, 1998). Instead, tests are used to detect the denotative similarity of word meanings, for example, "if the entity X in different situations can be called $W_1$, can it always be called $W_2$", and vice versa.

Thus, because of the above-mentioned differences (denotative tests, unique names of concepts), RuThes is closer to ontologies on an imaginary scale from lexical resources to formal ontologies than WordNet-like thesauri (Loukachevitch and Dobrov, 2014).

### 3.1 Relations in RuThes.

Different models of the knowledge description presuppose different sets of relations.

In RuThes, the relations are established only between concepts. The main class-subclass relation roughly corresponds to the relation of hyponym-hypernym in WordNet (Miller, 1998).

Also, RuThes has the part-whole relationship, but unlike WordNet, it is only established when the part always (or at least in the vast majority of cases) refers to the specified whole, i.e. cannot belong to a number of alternative wholes. This makes it possible to use the transitivity of the part-whole relations with greater reliability (Loukachevitch, Dobrov, 2014). There are some techniques allowing representation of part-whole relations in other cases.

When the above-mentioned conditions for establishing the part-whole relationship are imposed, a fairly broad interpretation of the part-whole relationship is adopted in RuThes:

- between physical objects (*storey – building*);
- between regions (*Europe – Eurasia*);
- between substances;
- between sets (*battalion – company*);
- between parts of the text (*strophe – poem*);
- between processes (*production cycle – industrial manufacturing*).

Also, the part-whole relations are established for connections between entities, one of which is internal, dependent on another (Guarino, 2009) such as: characteristics of an entity (*displacement – ship*); role in the process (*investor – investment*); participant in the field of activity is the sphere of activity (*industrial plant – industry*).

In addition, one of the main relations in RuThes is the relation of ontological dependence, which shows the dependence of the existence of one concept on another. An example of such an attitude is the relationship between the concepts *Tree – Forest*, where *Forest* is a dependent concept requiring the existence of the *Tree* concept.

The relation of the ontological dependence is denoted as directed association $asc_1 – asc_2$. In fact, this directed association represents a more formalized form of the association relations in traditional information-retrieval thesauri (Z39.19, 2005). Symmetric associations are also possible in only restricted number of cases.

Thus, the structure and the set of relations in the thesaurus RuThes are significantly different from the structure and relations of WordNet. It is also important to stress the differences in the properties of the relationships in the thesauri WordNet and RuThes. In WordNet, basically, only the transitivity of hyponym-hypernym relations is used. In RuThes, in addition to the transitivity of the class-subclass relationship, the following relations are also postulated:

- transitivity of the part-whole relations:

  $whole\ (c_1, c_2) \wedge whole\ (c_2, c_3) \rightarrow whole\ (c_1, c_3)$;

- inheritance of the whole relationship to subclasses:

  $class\ (c_1, c_2) \wedge whole\ (c_2, c_3) \rightarrow whole\ (c_1, c_3)$;

- inheritance of dependence association relations and symmetric association relations on types and parts:

$class\ (c_1, c_2) \wedge asc_1\ (c_2, c_3) \rightarrow asc_1\ (c_1, c_3);$

$class\ (c_1, c_2) \wedge asc\ (c_2, c_3) \rightarrow asc\ (c_1, c_3);$

$whole\ (c_1, c_2) \wedge asc_1\ (c_2, c_3) \rightarrow asc_1\ (c_1, c_3);$

$whole\ (c_1, c_2) \wedge asc\ (c_2, c_3) \rightarrow asc\ (c_1, c_3)$

Considering all possible relation paths existing between two thesaurus concepts $C_1$ and $C_2$, it was supposed that those paths that can be reduced to a single relation with the application of the above-mentioned rules of transitivity and inheritance indicate semantic relatedness between concepts $C_1$ and $C_2$, so called semantic paths. Word and phrases presented as thesaurus entries assigned to the concepts $C_1$ and $C_2$ are also considered semantically related even if the length of the path is quite large (five and more relations). Such defined semantic similarity between words and phrases included in RuThes is used for query expansion in information retrieval, thematic text representation (Loukachevitch and Alekseev, 2014), representation of categories in knowledge-based text categorization (Loukachevitch and Dobrov, 2015), and automatic word sense disambiguation.

The properties of the RuThes relations and defined paths were used to infer some types of relationships for RuWordNet.

### 3.2 Multiword Expressions in RuThes

Another issue, which is important in transformation of data from RuThes to RuWordNet, is the representation of multiword expressions (Loukachevitch and Lashevich, 2016).

The distinctive feature of RuThes is that it contains many multiword expressions. Experts are recommended to introduce new multiword expressions into RuThes if they can substantiate their decision with the necessity to represent the expression in the thesaurus. The expert should show that adding the expression to the thesaurus gives useful information that does not follow from the component structure of this expression. Such information is usually expressed in form of additional thesaurus relations (or their deliberate exclusion), which enriches the thesaurus knowledge.

In fact, we shift the often discussed question on compositionality vs. non-compositionality of a multiword expression to the more visible question of adding information to a thesaurus. The employed principles of introducing multiword expressions into RuThes can be subdivided as follows:

- absence of meaningful relations between an expression and senses of component words (idioms),
- synonym to own component word or its derivative (multisynonyms),
- additional relationships to other single words and multiword expressions.

In RuThes, multiword expressions that are synonymous its own component or its derivative are specially collected. The examples of such expressions include *политическая партия* (*political party*) – *партия* (*party*), the phrase is quite frequent in Russian as well as its translation in English. Another example is *компьютерная программа* (*computer program*) – *программа* (*program*). The example of a multisynonym to the component derivative is: *участвовать* (*participate*) – *принимать участие* (*take participation*).

In creating RuThes, the introduction of such multiword synonyms was especially encouraged, because the important feature of these expressions is that their components can be ambiguous, but the whole expression is often unambiguous. Thus, if the expression is known and described in a thesaurus there are no problems with disambiguation of its components and with the semantic interpretation of the whole expression. In fact, these expressions can improve the recognition of their own concepts.

In addition, the inclusion of such expressions in a synset often clarifies the sense of the synset. It is clear that introduction of these expressions does not require additional concepts.

Such multisynonyms are very common in the Russian language. Currently, the published version of RuThes – RuThes 2.0 (Loukachevitch et al., 2014) contains more than 13 thousand multiword synonyms.

Numerous examples of multisynonyms can be found also in English and can be met in WordNet. For example, *plant – industrial plant, platform – political platform, park – car park – parking lot.* But in RuThes, multisynonyms were specially searched and added.

RuThes also includes multiword expressions with so called *relational idiosyncrasy*, that is multiword expressions that look like compositional ones but they have specificity in relations with other single words and/or expressions, which usually means that these expressions denote some important concepts, entities or situations (Loukachevitch and Gerasimova, 2017).

For example, such phrase as *дорожное движение* (*road traffic*) seems to be compositional one, but it has hyponyms: *левостороннее движение* (*left-hand traffic*) and *правостороннее движение (right-hand traffic*): the existence of such hyponyms cannot be inferred from its component words.

Currently, all multiword expressions (54 thousands of 115 thousand entries) of RuThes-lite were transferred to RuWordNet. In such a way, it is possible to say that RuWordNet contains the maximal share of phrases in synsets among other WordNet-like resources. It means that the representation of phrases in RuWordNet requires special attention.

## 4 Creating Basic Structure of RuWordNet

In our opinion, 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 2.0 were subdivided into three parts of speech: nouns (single nouns, noun groups, or 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 7,636 verb synsets (Table 1).

This subdivision was based on the morpho-syntactic representation of RuThes-lite 2.0 text entries, which was fulfilled semi-automatically. Therefore, a small number of mistakes because of particle treatment (verbs or adjectives) or nominalized adjectives can appear. For example, Russian phrase *любитель подраться* (*=драчун*) (*brawler*, *scrapper*) was treated in this procedure as a verb group and was assigned to the verb synsets. Currently all found mistakes are corrected.

| Part of speech | Number of synsets | Number of unique entries | Number of senses |
|---|---|---|---|
| Noun | 29,296 | 68,695 | 77,153 |
| Verb | 7,634 | 26,356 | 35,067 |
| Adj. | 12,864 | 15,191 | 18,195 |

Table 1. Quantitative characteristics of synsets and entries in RuWordNet

The divided synsets were linked to each other with the relation of part-of-speech synonymy.

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 also established. By now, they had been generated semi-automatically for geographical objects.

The part-whole relations from RuThes were semi-automatically transferred and corrected according to traditions of WordNet-like resources. Now RuWordNet contains 3.5 thousand part-whole relations. The part-whole relations include the following subtypes:

- functional parts (*nostrils – nose*),
- ingredients (*additives – substance*),
- geographic parts (*Seville – Andalusia*),
- members (*monk – monastery*),
- dwellers (*Moscow citizen – Moscow*),
- temporal parts (*gambit – chess party*)
- inclusion of processes, activities (*industrial production – industrial cycle*)

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. For example, word *цветовой* (*colored*) is linked to such hyponyms as *красный* (*red*), *синий* (*blue)*, *зеленый* (*green)*, etc.

| Part of speech | Hyper-nyms | Inst-ance. | Holo-nyms | POS-syn. | Ant o-ny ms |
|---|---|---|---|---|---|
| Noun | 39,155 | 1863 | 10,010 | 18,179 | 454 |
| Verb | 10,304 | 0 | 0 | 7,143 | 20 |
| Adj. | 16,423 | 0 | 0 | 13,794 | 456 |

**Table 2**. Quantitative characteristics of basic relations in RuWordNet

Adjectives often have POS-synonymy links to nouns, but also can have POS-synonyms to verb synsets. For example, word *строительный* (*building* as an adjective) has two POS-synonymy relations: to the noun synset {*стройка, постройка, возведение,*

*сооружение*..} (*building* as a noun) and to the verb synset {*строить, построить, возводить* ...} (*to build*).

Antonymy relations are conceptual relations in RuWordNet, that means they link synsets, not single lexemes. They are introduced for all parts of speech, mainly for synsets denoting properties and states, for example:

- noun synset {*легкость, с легкостью, без труда, без затруднений*} (*easiness*) is antonymous to synset {*тяжесть, трудность*} (*difficulty)*,
- adjective synset {*легкий, легкий для выполнения, легкий для осуществления, нетрудный*} (*easy*) is antonymous to synset {*тяжкий, трудный, тяжелый, трудный для выполнения, нелегкий ...* } *(difficult*),
- verb synset {*не соответствовать действительности*} (*to be contrary to the fact)* is antonymous to synset {*соответствовать истине, соответствовать действительности*} (*to be in accordance with the truth)*.

The current numbers of basic relations described in RuWordNet are presented in Table 2.

## 5 Enrichment of Basic Relations of RuWordNet

Basic relations in the RuWordNet thesaurus were supplemented by several types of relations, including the relations of causation and entailment, the domain relation, the relations of word derivation and the relations between phrases and their components.

### 5.1 Causation and entailment

The relationships of entailment and causation were treated in the same way as in WordNet. The WordNet entailment relation is a relation between two verbs V1 and V2 that holds when the sentence "*Someone* $V_1$" logically entails "*Someone* $V_2$" and there is the temporal inclusion of event $V_1$ into $V_2$ or vice versa (Fellbaum, 1998). The causation relation can be also considered as a subtype of a general logical entailment relation but there is not temporal inclusion between corresponding situations (Fellbaum, 1998).

To automate the introduction of the relations of causation and entailment into RuWordNet, the RuThes directed associations between concepts containing verbs were extracted. This relation means in this case that the emergence of one situation (process, action) somehow requires the emergence of another situation (process, action).

The prepared lists of relations between verbs were checked out by linguists, resulting in the following relations:

- 97 relations of antonymy, denoting the opposite of what was before, for example, *откупорить* (*uncork) – закупорить (cork)*,
- 610 relations of causation, for example, *сажать* (*sit*) - *сесть* (*sit down*). This relation in RuWordNet often connects the synsets corresponding to the reflexive forms of the verbs, for example, the synset *купать*, *выкупать*, *докупать*, *искупать* (*give a bath*) is the cause of *купаться, выкупаться, искупаться, покупаться* (*to bathe, cleanse own body*).
- 943 entailment relationships, for example, the synset *сниться (to dream*) is related by the entailment relation with synset *спать*, *поспать*, *почивать (to sleep*) because if someone dreams something, then this someone is sleeping.

### 5.2 Domain relations

Since relations in such thesauri as WordNet are mostly generic (hyponym-hypernym), there exists a so-called "tennis problem" (Miller, 1998), which is that synsets from the same domain (for example, related to tennis: *tennis player, racket, court*) are very far from each other in the WordNet hierarchy.

To solve this problem in part, a hierarchical system of domains (domains)[3] has been proposed, and WordNet synsets were semi-automatically assigned to one or more domains (Magnini, Pianta, 2000; Bentivogli et al., 2004). This domain system is now partially transferred to RuWordNet.

The mechanism of introducing domains for the RuWordNet synsets was as follows. The existing domain system for Princeton WordNet was taken. First, the domain list was refined: the subject areas that were not presented in the RuWordNet thesaurus were removed (i.e. *Heraldry*), and several new domains were added. For example, domain labels corresponding to world religions and some confessions were introduced. Currently, RuWordNet has 156 domains.

The domains labels can be considered as a list of categories for a knowledge-based categoriza-

[3] http://wndomains.fbk.eu/

tion system. RuThes has a special interface for linking categories with thesaurus concepts and hierarchies.

Each domain was linked to one or more "supporting" concepts of the RuThes thesaurus. Using the RuThes relation properties, the list of supporting concepts was expanded by lower-level concepts (subclasses, parts, associations). This can be done, because in RuThes the relation to the sphere of activity is one of the types of the part-whole relationship, and therefore it is explicitly indicated in the thesaurus.

The generated list of concepts for each domain was looked through and cleaned by experts. Also, for each domain, a noun synonym of RuWordNet was assigned as the domain title.

As a result, a chain of relations has been created:

(1) RuWordNet synsets,

(2) Initial concepts of the RuThes thesaurus for these synsets,

(3) Domain labels presented as categories over RuThes concepts,

(4) RuWordNet synsets, assigned as a label to each subject domain.

Such a chain makes it possible to introduce direct domain relations between RuWordNet synsets: (1) -> (4).

For example, domain "Art" is described as RuThes concept *Art* with full expansion, which adds to the Art domain all hyponyms, parts, dependent concepts obtained by logical inference using the properties of transitivity and inheritance (Section 3.1). As a result, "Art" concepts comprise more than 700 RuThes concepts, including *Jazzman, Piece of painting, Harp*, etc. Then RuWordNet synsets originated from these RuThes concepts were also assigned to the Art domain.

### 5.3 Derivational relations

For RuWordNet, the derivational relations were also introduced (Leseva et al., 2015; Pala and Hlaváčková, 2007, Piasecki, et al, 2012). These relations are lexical, that is established between lexical entries. At the moment, these relations are established for those words that have the same beginning of the word (without prefixes).

The derivation relations were established between words if two conditions were fulfilled:

- the words have the same beginnings,
- these words refer to concepts that either have a direct relationship in the RuThes thesaurus or the relationship can be derived from the properties of transitivity and inheritance established in RuThes.

For example, for the word *аренда* (*lease*), the following words with the same root are indicated: *арендатор* (*lessee*), *арендаторский* (*lessee* as an adjective), *арендователь* (*lessee*), *арендаториша* (*lessee-woman), арендный* (*lease* as an adjective), *арендование* (*leasing*), арендовать (*to lease*), *арендодатель* (*lease-holder*). Such relations allow us to present semantic relations between words for which there is no other suitable relationships in RuWordNet.

### 5.4 Relations between phrases and its components

According to the accepted rules for the RuThes thesaurus, experts try to find all possible words and phrases that can express a specific concept (Loukachevitch and Lashevich, 2016). In addition, as described in subsection 3.2, a new concept can be introduced if a phrase carries information that does not follow from the meanings of the word-components of this phrase. For example, RuThes contains the concept *Increase of prices*, which have an important relation to the concept of *Inflation*. Text entries of the concept in RuThes comprise a variety of phrases as: *price growth, increase prices, price increases*, etc.

This decision in RuThes is supported with the existing system of relations. For example, we can easily describe relations between concepts *Price, Increase of prices* and *Inflation* using directed associations.

| Type of relation between word and phrase | Number of relations |
|---|---|
| Phrase and its component are in the same synset (*political party – party*) | 13,367 |
| Pos-synonym relations (*participate – take participation*) | 6,285 |
| Other relations from RuWordNet | 16,279 |
| Direct RuThes relations, not included in RuWordNet | 15,677 |
| Relations inferred using the RuThes relations properties | 12,513 |

Table 3. Quantitative characteristics of the relationships between phrases and their components in RuWordNet

All these solutions lead to a large number of multiword expressions in RuThes. When RuWordNet has been generated, the phrases were also transferred to it from RuThes. However, the RuWordNet relationship system is different, and for a large number of compositional phrases, the relationship between the phrase and its component words can be lost, which can negatively affect the use of the RuWordNet thesaurus in natural language processing. Therefore, in RuWordNet additional types of relations have been introduced: for the phrase (*has_component*) and for individual words that are phrase components (*component_for*).

These relations were obtained automatically on the basis of direct relations in the thesaurus RuThes, and also on the basis of a logical inference on the relation properties (Section 3.1). Table 3 shows the quantitative results for the established relations between phrases and its components in RuWordNet.

## Conclusion

In the paper, we presented a new Russian wordnet, RuWordNet, which was obtained by semi-automatic transformation of the existing Russian thesaurus RuThes. At the first step, the basic structure of wordnets was reproduced: synsets' hierarchies for each part of speech and the basic set of relations between synsets (hyponym-hypernym, part-whole, antonyms).

At the second stage, we added causation, entailment and domain relations between synsets. Also, derivation relations were described for single words and component structure for phrases included in RuWordNet.

It can be seen that RuThes relations are unusual for wordnet-like resources but they give the possibility:

- to introduce a multiword expression into the thesaurus if it gives new information,
- infer domain labels because in RuThes the domain relation is a subtype of the part-whole relation,
- infer derivation relations between lexical entries using the RuThes relation properties.

**Acknowledgments**
This work is partially supported by Russian Scientific Foundation, according to the research project No. 16-18-020.

## References

Saima Aman and Stan Szpakowicz. 2008. Using Roget's Thesaurus for Fine-grained Emotion Recognition, *Proceedings of IJCNL-2008*: 312-318.

Irina Azarowa. 2008. RussNet as a Computer Lexicon for Russian, *Proceedings of the Intelligent Information systems IIS-2008*: 341-350.

Irina Azarova, Pavel Braslavski, Viktor Zakharov, Yuri Kiselev, Dmitrii Ustalov and Maria Khohlova. 2016. Integration of thesauri RussNet и YARN. Proceedings of "Iternet and Modern Society" conference IMS-2016 (in Russian).

Valentina Balkova, Andrey Suhonogov, and Sergey Yablonsky. 2008. Some Issues in the Construction of a Russian WordNet Grid. *In Proceedings of the Forth International WordNet Conference*, Szeged, Hungary:44-55.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, B., and Emanueke Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources:* 101-108.

Francis Bond, and Paik Kyonghee. 2012. A survey of wordnets and their licenses. *Proceedings of Global Wornet Conference GWC-2012*: 64-71.

Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources*, Odense.

Lars Borin, Markus Forsberg and Lennart Lonngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Lars Borin, Jens Allwood, and Gerard de Melo. 2014 .Bring vs. MTRoget: Evaluating automatic thesaurus translation. *Proceedings of LREC-2014*, Reykjavık, ELRA: 2115–2121.

Pavel Braslavski, Dmitrii Ustalov and Mikhail Mukhin. 2014, A Spinning Wheel for Yarn: User Interface for a Crowdsourced Thesaurus, *Proceedings of EACL-2014*, Gothenberg, Sweden: 101–104.

Boris Dobrov and Natalia Loukachevitch. 2006. In Development of Linguistic Ontology on Natural Sciences and Technology.In *Proceedings of LREC-2006.*

Boris Dobrov and Andrey Pavlov. 2010. A Basic line for news clusterization methods evaluation. *Proceedings of RCDL-2010*: 287-295.

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

Alfiya Galieva, Olga Nevzorova and Dilyara Yakubova . .2017. Russian-Tatar Socio-Political Thesaurus: Methodology, Challenges, the Status of the Project. In Proceedings of Recent Advances in Natural Language Processing Conference (RANLP-2017): 245-252

Ilia Gelfenbeyn, Artem Goncharuk, Vlad Lehelt, Anton Lipatov, and Viktor Shilo. 2003. Automatic translation of WordNet semantic network to Russian language. *In Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2003.*

Derek Gross and Katherine Miller. 1990. Adjectives in WordNet, *International Journal of Lexicography*, *3*(4):.265-277.

Nicola Guarino. 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, Berlin and Heidelberg, Germany: Springer-Verlag: 52–67.

Mario Jarmasz and Stan Szpakowicz. 2004. Roget's thesaurus and semantic similarity, *Recent Advances in Natural Language Processing III: Selected Papers from RANLP*, 2004: 111-120.

Alistair Kennedy, and Stan Szpakowicz. 2008. Evaluating Roget's Thesauri, *Proceedings of ACL-2008*: 416-424.

Mikhail Khodak, Andrej Risteski, Christiane Fellbaum, and Sanjeev Arora. 2017. Automated WordNet Construction Using Word Embeddings. *Proceedings of SENSE 2017*: 12-23.

Claudia Kunze and Lothar Lemnitzer. 2010. Lexical-Semantic and Conceptual relations in GermaNet. *In Storjohann P (ed) Lexical-semantic relations: Theoretical and practical perspectives,* 28:163-183.

Svetlozara Leseva, Maria Todorova, Tsvetlana Dimitrova, Borislav Rizov, Ivelina Stoyanova, Svetla Koeva. 2015. Automatic classification of wordnet morphosemantic relations. In *The 5th Workshop on Balto-Slavic Natural Language Processing*: 59-64.

Krister Lindén and Jyrki Niemi. 2014. Is it possible to create a very large wordnet in 100 days? An evaluation, *Language resources and evaluation*, 48.2: 191-201.

Natalia Loukachevitch and Boris Dobrov. 2002. Development and Use of Thesaurus of Russian Language RuThes. *Proceedings of workshop on WordNet Structures and Standartisation, and How These Affect WordNet Applications and Evaluation*. *LREC-2002*: 65-70.

Natalia Loukachevitch and Boris Dobrov. 2014. RuThes Linguistic Ontology vs. Russian Wordnets. *Proceedings of Global WordNet Conference GWC-2014, Tartu*: 154-162.

Natalia Loukachevitch, Boris Dobrov and Ilia Chetviorkin. 2014. RuThes-Lite, a publicly available version of thesaurus of Russian language RuThes. In *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2014*, 340-350.

Natalia Loukachevitch and Aleksey Alekseev. 2014. Summarizing News Clusters on the Basis of Thematic Chains. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Natalia Loukachevitch and Boris Dobrov. 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.

Natalia Loukachevitch and German Lashevich. 2016. Multiword expressions in Russian thesauri RuThes and RuWordNet. In *Artificial Intelligence and Natural Language Conference* (*AINL-2016*), IEEE: 1-6.

Natalia Loukachevitch, Alexander Shevelev and Valeria Mozharova. 2017. Testing Features and Methods in Russian Paraphrasing Task. In *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog-2017*, V.1:. 135-146.

Natalia Loukachevitch and Anastasia Gerasimova. 2017. Human Associations Help to Detect Conventionalized Multiword Expressions. *arXiv preprint arXiv:1709.03925*.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of Language Resources and Evaluation Conference LREC-2000*: 1413-1418.

Inderjeet Mani, I., Klein, G., House, D., Hirschman, L., Firmin, T., and Sundheim, B. (2002). SUMMAC: a text summarization evaluation. *Natural Language Engineering*, 8 (1), 43-68.

Marek Maziarz, Stanoslaw Szpakowicz and Maciej Piasecki. 2012. Semantic relations among adjectives in Polish WordNet 2.0: a new relation set, discussion and evaluation. *Cognitive Studies*, *12*:,149-179.

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

George Miller and Florentina Hristea. 2006. WordNet Nouns: Classes and Instances. *Journal of Computational linguistics*, 32(1):1-3.

Michael Nokel and Natalia Loukachevitch. 2016. Accounting ngrams and multi-word terms can improve topic models. In *Proceedings of 12th Workshop on Multiword Expressions, ACL 2016:* 44-49.

Karel Pala and Dana Hlaváčková. 2007. Derivational relations in czech wordnet. *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*. Association for Computational Linguistics.

Bolette Pedersen, Sanni Nimb, Jorg Asmussen, Nicolai Sørensen, Lars Trap-Jensen L and Henrik Lorentzen. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary, Language resources and evaluation, 43(3): 269-299.

Maciej Piasecki, Radoslaw Ramocki, and Marek Maziarz. 2012. Automated generation of derivative relations in the Wordnet expansion perspective. In *Proceedings of the 6th Global Wordnet Conference GWC 2012*: 273–280.

Piek Vossen. 1998. Introduction to EuroWordNet. In *EuroWordNet: A multilingual database with lexical semantic networks*, Springer Netherlands: 1-17.

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