Linking Russian Wordnet RuWordNet to WordNet

Natalia Loukachevitch
Lomonosov Moscow State University
Moscow, Russia
Louk_nat@mail.ru

Anastasia Gerasimova
Lomonosov Moscow State University
Moscow, Russia
anastasiagerasimova432@gmail.com

Abstract

In this paper we consider the linking procedure of Russian wordnet (RuWordNet) to Wordnet. The specificity of the procedure in our case is based on the fact that a lot of bilingual (Russian and English) lexical data have been gathered in another Russian thesaurus RuThes, which has a different structure than WordNet. Previously, RuThes has been semi-automatically transformed into RuWordNet, having the WordNet-like structure. Now, the RuThes English data are utilized to establish matching from the RuWordNet synsets to the WordNet synsets.

1 Introduction

The Princeton WordNet thesaurus (Fellbaum, 1998, Miller, 1998) created for the English language is one of the most popular linguistic resources used in natural language processing. In many countries their own projects on creating WordNet-like resources (wordnets) for national languages have been initiated (Vossen, 1998).

The Open Multilingual WordNet project is currently being developed (Bond and Paik, 2012; Bond and Foster, 2013; Rudnicka et al., 2017). The goal of the project is to link together the existing wordnets created for different languages with an open license[1]. To connect a new language to the project, it is necessary to associate synsets of this language with WordNet synsets and present the data in the required format.

Sources of links of a specific wordnet to English synsets of Princeton WordNet can be different (Vossen, 1998; Piante et al., 2002). Some wordnets have been developed with semi-automatic translation of Princeton WordNet synsets, and therefore these links exist from the beginning. The creators of the Finnish wordnet (FiWN) translated 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). Other wordnets are developed from scratch using own-language text corpora and dictionaries (Rudnicka et al., 2017). In such cases, their linking to WordNet synsets should be organized as a special procedure based on bilingual dictionaries and expert verification.

In the current study, we describe another way of aligning the Russian wordnet (RuWordNet) and WordNet synsets. RuWordNet was semi-automatically generated from another Russian thesaurus RuThes, which is being developed for more than 20 years (Loukachevtich et al., 2018; Kirillovich et al., 2017). For bilingual text processing, the RuThes concepts also have English representation. This English part of the RuThes thesaurus has been collected from various sources, including several text collections (news articles, European Community documents, etc.), English and Russian-English dictionaries, and others. Currently, the RuThes concepts have more than 140 thousand English text entries. In the paper we describe the process of linking RuWordNet with WordNet, which exploits the previously gathered bilingual data in RuThes.

The paper is structured as follows. In Section 2 we consider related work. Section 3 describes RuWordNet thesaurus and its source - RuThes thesaurus, including representation of bilingual Russian-English lexical units and phrases. Also the general scheme of links. In Section 4 we consider the general scheme of linking RuWordNet and WordNet using RuThes bilingual data. Section 5 presents two main steps of linking RuWordNet and WordNet: automated linking through RuThes bilingual information and manual linking of WordNet core concepts.

[1] http: // compling.hss.ntu.edu.sg/omw/

2 Related Work

For the first time, the idea of linking wordnets was proclaimed in EuroWordNet project (Vossen, 1998). In order to establish communication between different languages, the synsets of each wordnet should refer to the so-called interlingual index (ILI), for which the Princeton WordNet synsets were used. The index is an unordered list of synsets with glosses. To accurately describe the correspondence of specific synsets of each language and overcoming lexical gaps that may arise in a particular language, several different equivalence relations from synsets of a specific language to the ILI index were proposed: synonym, near-synonym, hyperonym, hyponym.

Christea et al. (2004) list the main problems of linking English-language WordNet and another wordnet using Romanian wordnet (Tufiş et al., 2013) as an example. The first type of difficulties is related to the fact that potential matches in WordNet correspond to several synsets denoting similar senses, and the explanations of synsets are very similar. Additional analysis is needed to choose the most appropriate synset.

The second type of problems is associated with the absence of lexicalized means of naming a concept denoted by the English synset. In such cases, an additional synset is introduced into the Romanian wordnet, which contains a non-lexicalized expression. The next type of problems stems from the fact that the word sense system in the English WordNet is more fractional than in the Romanian wordnet. In such cases, new senses were entered into the Romanian wordnet.

Linking between Polish wordnet (plWordNet) and WordNet was performed in 2012 (Rudnicka et al., 2012). To establish links, the following set of interlingual (I) relationships was used: I-synonymy, I-hyponymy, I-hyperonymy, I-meronymy, I-holonymy, I-quasi-synonymy (near synonymy), I- inter-register synonymy. The latter relation is established when the synsets in Polish and English have the same meaning, but refer to different language registers. The matching between the Polish and English synsets was performed manually. In the process of searching for equivalents, inaccurate descriptions of Polish word senses could be corrected.

Maziarz et al. (2013) provide quantitative characteristics of the established relations: the I-hyponymy relation was the most frequent link between synsets of WordNet and plWordNet. This can be explained by the existence of a large number of lexical and cultural lacunae, greater lexicalization of the category of gender in the Polish language (for example, for the names of roles, posts of people), the use of diminutive names in Polish, etc.

3 RuWordNet Thesaurus

The Russian wordnet RuWordNet (Loukachevitch et al., 2016; Loukachevitch et al., 2018) has been created on the basis of another Russian thesaurus RuThes in 2016 (Loukachevitch, Dobrov, 2002).

Main units of RuThes are concepts, each concept has a monosemous and clear name and the set of text entries that convey the corresponding concept in texts. The text entries of a concept can include single words of different parts of speech, multiword expressions and also compositional phrases, with the same meaning. To represent bilingual data, the RuThes concept has the English name of concept and the set of English text entries with the same variety of text entries.

To create RuWordNet, the RuThes data were transformed: the concepts were subdivided to part-of-speech-related synsets and traditional WordNet-like relations were established between the synsets. Table 1 presents the quantitative characteristics of synsets and language units in RuWordNet.

Further we consider the organization of English part in the RuThes because we use these data for linking RuWordNet and WordNet.

Part of speech	Number of synsets	Number of unique Russian 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 the synsets and Russian entries in RuWordNet

3.1 RuThes as a Bilingual Resource

RuThes is a linguistic ontology presented as a hierarchy of concepts. Each concept has a unique name in Russian and in English (if existing). A concept is associated with a set of Russian text entries and English text entries.

Text entries of the same concepts in both languages can include single words of different parts of speech, multiword expressions, and compositional phrases that can express this con-

cept. Current volume of RuThes is more than 60 thousand concepts, 200 thousand Russian text entries and 146 thousand English text entries.

The English text entries were collected for many years from several sources, including bilingual dictionaries, analysis of English documents in various projects, such as knowledge-based text categorization.

During last years, each new concept introduced into RuThes is provided with the English name and English text entries, if they exist. These English translations are specially searched in bilingual resources or translated with online-translation services. Then all English variants are verified on Internet-pages to check if they really exist and express the intended senses, because any found translations can be incorrect.

Besides direct translations, also cross-category synonyms are added as text entries, for example, adjective or verb derivations expressing the same concept. Additionally, multiword phrases expressing the same concept are searched for and introduced, because for various applications it is important to match a thesaurus concept in texts using its variant forms.

For example, for concept *ПРОМЫШЛЕННОСТЬ (promyshlennost')/ INDUSTRY* the following English text entries have been introduced: *industry*, *industrial*, *industrial sphere*, *sphere of industry*. From this example, the importance of adding such multiword variants can be seen: they are unambiguous, but their components have several senses.

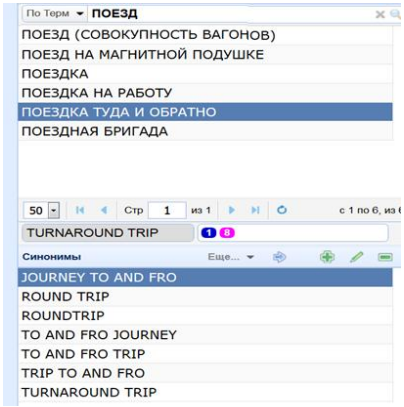

Figure 1. English text entries for the RuThes concept *ПОЕЗДКА ТУДА И ОБРАТНО* (*TURNAROUND TRIP*)

Figure 1 shows English variants collected for the RuThes concept *ПОЕЗДКА ТУДА И ОБРАТНО* (*TURNAROUND TRIP*). It could be noted that corresponding synset in WordNet contains only the *round trip* lexical entry.

Figure 2 demonstrates English text entries for the RuThes concept *ПОЕЗДКА НА РАБОТУ* (*COMMUTE TO WORK*). In WordNet word *commute* has 1 noun sense and 5 verb senses, which means that this word can be quite difficult for word sense disambiguation. But when we introduce unambiguous variant phrases *commute for work* and *commute to work*, we provide reliable way to detect this concept in texts because these phrases are quite frequent according to Google (*commute for work* – 143 thousand pages, *commute to work* – 12 mln. pages).

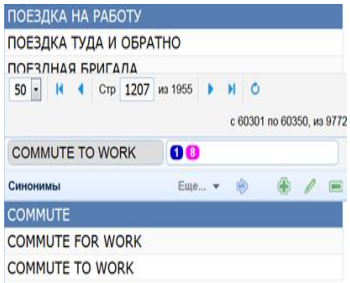

Figure 2. English text entries for the RuThes concept *ПОЕЗДКА НА РАБОТУ* (*COMMUTE TO WORK*)

RuThes is a Russian-oriented resource. In such cases when a single Russian word corresponding to an English word sense is absent, the following solutions can be made:

- If the sense can be expressed with an existing Russian phrase (multiword expression or a compositional phrase) then an additional concept can be introduced,
- in other cases, such English word can be attached to the closest RuThes concept. For example, English word *watch* (portable timepiece) is linked to the RuThes concept *ЧАСЫ* (*TIMEPIECE*) (Figure 3)

On Figure 3 the upper left form contains a list of concepts with "часы" substring. The lower left form shows text entries for the highlighted concept. In the middle between these forms, the English concept name (*TIMEPIECE*) can be seen. The right upper form presents the relations of the highlighted concept.

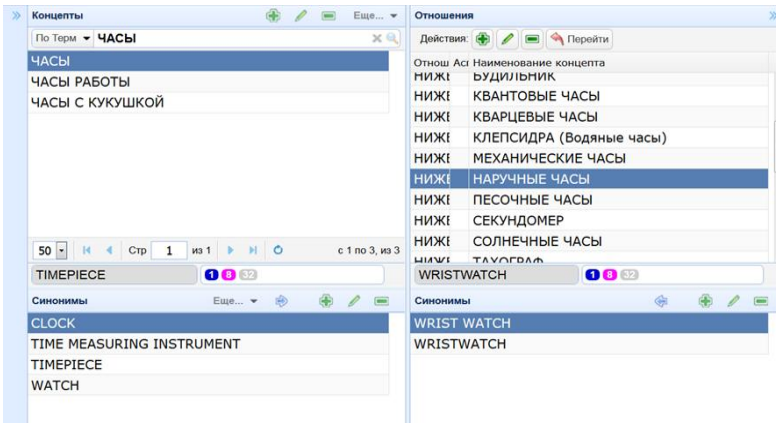

Figure 3. The differences in conceptualization of timepieces in Russian and English: there is no Russian word for English *watch*, as a portable timepiece

The low right form of Fig. 3 describes text entries of the highlighted concept *НАРУЧНЫЕ ЧАСЫ* (*WRIST WATCH*).

4 General Scheme of Linking RuWordNet to WordNet

The synsets of RuWordNet contain reference links to RuThes concepts from which these synsets were generated. Therefore English text entries collected in the English part of RuThes now can be used for matching RuWordNet and WordNet synsets.

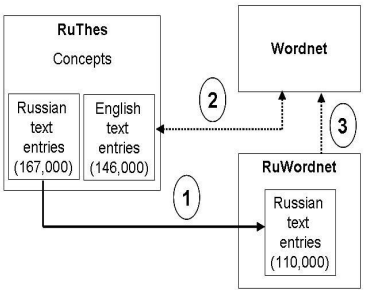

Figure 4. The scheme of linking RuWordNet to WordNet through the RuThes concepts with English text entries

Figure 1 shows the connections between the resources. Initially, thesaurus RuThes has been created. Most concepts of RuThes have Russian and English names and Russian and English text entries. Then the Russian part of RuThes was semi-automatically transformed to the WordNet-like thesaurus RuWordNet (**link 1**). Currently, we are semi-automatically creating links between the English part of RuThes and the WordNet synsets (**link 2**). From these two procedures, we obtain links from the RuWordNet synsets to the WordNet synsets (**link 3**).

5 Linking Procedure

The process of linking of WordNet and RuWordNet synsets includes two parts:

- Automatic matching the RuThes English entries with the WordNet units with further validation by experts and the transfer of the Russian established link from RuThes to RuWordNet, which has direct correspondence with RuThes,
- Analysis of the core wordnet synsets (Boyd-Graber et al., 2006), which are considered to be frequent and most salient. The task of the analysis is to check if the English-Russian links were established, or some corrections are needed, or the link cannot be established because of the absence of proper lexicalization in Russian.

Currently, I-S (inter-language synonym) and I-NS (inter-language quasi-synonym) are established between WordNet and RuWordNet synsets (through RuThes concepts). The relationship of interlanguage synonymy is established if the synset and concept have very close sets of denotations, but there are some features of the

word meanings that are different in the two languages

In subsections we consider these two procedures and their results.

5.1 Linking translated RuThes Concepts

English text entries of RuThes were automatically matched with WordNet entries. Table 2 shows the main types of situations that occurred as the result of the performed matching for nouns. Let us consider some examples for each type of linking of the RuThes concepts and WordNet synsets.

Type 1.1. (one-to-one) links are usually represented by the concepts of certain domains, for example, chemistry (*hydrogen, helium*), finance (*credit system, central bank*), politics (*communist party, iron curtain*), medicine (*thrombophlebitis, bronchial asthma*), geographical names (*Minsk, White sea*), names of animals and plants, etc.

Types of matching between RuThes concepts and WordNet noun synsets	Number of RuThes concepts
1. RuThes concept has only single English text entry, among them:	**9,629**
1.1. One-to-one matching with WordNet synset	1,373
1.2. One-to-many matching with WordNet synsets	4 ,935
1.3. No matching with WordNet synsets	3,803
2. RuThes concept has several English text entries, among them:	**19,715**
2.1. Only one English text entry has single matching with a WordNet synset	4,343
2.2. Several English text entries correspond to monosemous WordNet units	3,344
2.2.1. Several English text entries mainly match with one of the WordNet synsets	1,611
2.3. Several text entries and all their matches with WordNet are ambiguous	4,425
2.4. Several English text entries but none of them matches with WordNet units	5,589

Table 2. The quantitative results of automatic matching English text entries in RuThes and the WordNet synsets

As an example of the **1.2 type of links,** the word *energy* can be considered, which is the only option in RuThes for the concept *ENERGY* as a physical characteristic, and also corresponds to the concept *HUMAN ENERGY* in the group of synonyms (*energy, human energy, life energy, vigor, vigor*).

In WordNet, the word *energy* is included into 7 synsets of nouns, one of which obviously corresponds to the physical meaning of the word energy (as in RuThes). One of the senses in WordNet corresponds to energy as a specific state of mind, enthusiasm. This sense clearly exists in Russian, but is absent in RuThes, and should be added.

Therewith, the word *energy* is attributed by the authors of WordNet to the synset: *Department of Energy, DOE (Department of Energy, United States; created in 1977)*. In RuThes, there is a similar entity, called *Министерство топлива и энергетики* (*Ministry of Fuel and Energy*) with the translations: *Department of Energy, Energy department,* etc, but the text entry *energy* is absent. In this case, the RuThes concept and the WordNet synset will be matched by other text entries (**type of comparison 2.3.).**

Some of the RuThes concepts and WordNet synsets cannot be matched, when a WordNet synset includes only single words, but in RuThes the related concept is linked only with phrases as text entries. For example, for the RuThes concept *ЗОЛОТОЙ ЦВЕТ* (*golden color*) there is a direct analogue in WordNet, namely synset: *(n) amber, gold (a deep yellow color)*. However, RuThes contains only English noun phrases as text entries: *golden color, gold color, golden colour, gold colour*.

The above-mentioned example of the synset *amber, gold* also demonstrates another problem, which arises from the comparison of two thesauri for different languages, namely the differences in conceptualization, i.e. what exactly is considered in each resource to be the same concepts, and what is considered to be different. Conceptualization may be erroneous in one of the resources. In some cases it may be not clear enough how it is better to divide words into synsets (attributed to concepts).

The unified synset *amber, gold* in WordNet means that the concepts of golden and amber colors are united in WordNet, while in RuThes they have different concepts. Description and comparison of different colors and their shades is a difficult task. However, the existing systems for presenting colors on the html pages of the Internet, for example, distinguish between amber and gold colors, matching code FFD700 to the gold color, and code FFBF00 to the amber color, that is, the RuThes presentation is more correct.

It is possible to find examples of another kind, when two synsets of WordNet correspond to a single RuThes concept. For example, in RuThes there is the concept *АТОМНАЯ ЭНЕРГИЯ* (*atomic energy*), the text entries for which in Russian are the phrases *атомная энергия* (*atomic energy*) and *ядерная энергия* (*nuclear energy*), and in English the name of this concept is formulated as *NUCLEAR ENERGY*, and the following phrases are listed as text entries*: atomic energy*, *atomic power*, *nuclear energy*, *nuclear pow*er.

In WordNet, two synsets correspond to this single RuThes concept: 1) atomic energy, nuclear energy (energy released by a nuclear reaction); 2) atomic power, for civilian use. In the second synset, *atomic power* is considered as a function of the atomic energy from the first synset, namely the use in power engineering. However, it seems that the same treatment of this sense cannot be reproduced in Russian.

Another example of the differences in conceptualization is related to the concept of *clock*. There are three basic concepts in WordNet: *time piece*, *timekeeper*, *horologe* and its two hyponyms: *clock* (a timepiece that shows the time of day) and *watch*, *ticker* (a small portable timepiece), including wrist or pocket watches.

Wikipedia shows a different type of conceptualization of these concepts for the English language, when *clock* and *timepiece* are united into one article, and the watch has another article. In RuThes, there is one concept of *ЧАСЫ* (*Timepiece)*, with English-language translations: *clock*, *watch*, *timepiece*, and various subspecies of clocks, since in Russian there is no more general concept corresponding to the dimension of time than *часы* (clock), nor individual words that correspond to small, "portable" clocks.

Thus, it can be seen that the comparison between semantic systems of different resources reveals flaws (repetition of sense, lack of senses) in one of the descriptions or different conceptualizations. Therefore, it is hardly worth setting the task of complete linking of all concepts (synsets).

It can be seen from the Table 2 that the published version of RuThes contains about 9 thousand concepts (of 31 thousand concepts), which have English text entries but no matching with WordNet noun synsets (**Types 1.3** and **2.4**). These concepts include:

- Russian and near-to Russia geographic names (about 1300 concepts),
- concepts having only verbs or adjectives as text entries,
- Russia-specific cultural and social concepts: *gzhel* (Russian style of blue and white ceramics), *sopka* (specific hills in Siberia), *kalach* (Eastern European bread), *kissel* (viscous fruit dish), *gorodki* (ancient Russian folk sport),etc.,
- concepts based on multiword expressions, which are absent in WordNet.

The direct matching of RuThes concepts and WordNet synsets, utilizing unambiguous and the most frequent correspondences (with post-editing), gave the following numbers of the established links between RuWordNet and WordNet synsets:

- 8,608 from 29,296 noun synsets,
- 996 from 7,634 verb synsets,
- 2,100 from 12,864 adjective synsets.

5.2 Translating Core Concepts

Additionally to the above-described matching to WordNet based on the RuThes English text entries, the independent examination of the WordNet core synsets is necessary because some English words can be absent in the English counterpart of the RuThes thesaurus. In this case, a professional linguist searches for each WordNet core synset direct link to a RuWordNet synset using both English text entries from RuThes and also any additional resources.

Currently, we have 90% of synonym and near-synonym links for the WordNet core concepts with the RuWordNet synsets, and it seems a very high level for the resources, which have been developed independently. About 400 new RuWordNet synsets have been proposed to introduction.

Table 3 shows statistics on established relations between RuWordNet and WordNet synsets for core synsets.

Part of Speech	Number of core concepts	Percent of established links (%)
Nouns	3300	90.3
Adjectives	698	85.0
Verbs	999	94.0
Total	4997	90.0

Table 3. Statistics on established relations between the RuWordNet and WordNet synsets for the core synsets

Some examples of core WordNet noun synsets for which the correspondece in RuWordNet are metonymic transfer of source senses:

- (n) village, small town, settlement (a community of people smaller than a town)
- (n) university (the body of faculty and students at a university)
- (n) manner of speaking, speech, delivery (your characteristic style or manner of expressing yourself orally)

Other examples of absent noun links are quite diverse:

- (n) style (editorial directions to be followed in spelling and punctuation and capitalization and typographical display)
- (n) survivor (one who outlives another) "*he left his farm to his survivors*"
- (n) search (an investigation seeking answers) "*a thorough search of the ledgers revealed nothing*

For adjectives, the most frequent problems of linking between two resources is the absence of an adjective form for a specific concept, which can be expressed with a particle (that is a verb form) in Russian. For example, the following "core" adjectives senses are absent in Russian:

- absent – *отсутствующий* (otsutstvuyushchiy),
- *afraid* – *испуганный* (*ispuganyy*),
- *asleep* – *спящий* (*spyashchiy*).

The main reason of absense of verbal links is due that such senses are expressed only with *light verb+noun* constructions in Russian:

- [cast]: select for a play or movie,
- [cater] supply food ready to eat,
- [demonstrate] march, march in protest.

6 Conclusion

In this paper we have considered the procedure for linking Russian wordnet (RuWordNet) to WordNet. The specificity of the procedure is based on the fact that a lot of bilingual (Russian and English) lexical data have been gathered in another Russian thesaurus RuThes, which has the structure different from WordNet. At first, Russian wordnet was semi-automatically generated from RuThes. Now, the RuThes English data are utilized to establish matching from the RuWordNet synsets to the WordNet synsets (through RuThes concepts).

Additionally, the WordNet core concepts are manually looked through to establish direct relations between RuWordNet and WordNet. Currently, 90% of the core Wordnet synsets are provided with links to RuWordNet, which is quite a large percentage for the independently developed resources.

Acknowledgments

The reported study was funded by RFBR according to the research project N 18-00-01226 (18-00-01240).

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*: 1352-1362

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

Jordan Boyd-Graber, Christiane, Fellbaum, D. Osherson, and R. Schapire. 2006. Adding dense, weighted connections to WordNet.' In: *Proceedings of the Third Global WordNet Meeting, GWC-2006*.

Dan Cristea, Catalin Mihaila, Corina Forascu, Diana Trandabat, Maria Husarciuc, Gabriela Haja, and Oana Postolache. 2004. Mapping Princeton WordNet synsets onto Romanian WordNet synsets. *Romanian Journal of Information Science and Technology*, *7*(1-2): 125-145.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Alexander Kirillovich, Olga Nevzorova, Emil Gimadiev, and Natalia Loukachevitch. RuThes Cloud: Towards a Multilevel Linguistic Linked Open Data Resource for Russian. In: P. Różewski and C. Lange (eds.) Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web (KESW 2017). Communications in Computer and Information Science, vol. 786, pp. 38-52. Springer (2017)

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 Lan-

guage RuThes. *Proceedings of workshop on WordNet Structures and Standartisation, and How These Affect WordNet Applications and Evaluation, LREC 2002*: 65-70.

Natalia Loukachevitch, German Lashevich, Anastasia Gerasimova, Vladimir Ivanov, and Boris Dobrov. 2016. Creating Russian WordNet by Conversion. In *Proceedings of Conference on Computational linguistics and Intellectual technologies Dialog-2016*: 405-415

Natalia Loukachevitch, German Lashevich, and Boris Dobrov. 2018. Comparing Two Thesaurus Representations for Russian. *Proceedings of Global WordNet Conference GWC-2018*: 35-44.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013. Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with WordNet. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*: 443-452.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. *Proceedings of the First International Conference on Global WordNet*: 293-302.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. 2012. A strategy of Mapping Polish WordNet onto Princeton WordNet. *In Proceedings of COLING 2012*: 1039-1048.

Ewa Rudnicka, Maciej Piasecki, Piotrowski, T., L. Grabowski, and Francis Bond. 2017. Mapping wordnets from the perspective of inter-lingual equivalence. *Cognitive Studies| Études cognitives*, (17).

Dan Tufiş, Verginica Mititelu, Dan Ştefănescu, and Radu Ion. 2013. The Romanian wordnet in a nutshell. *Language resources and evaluation*, *47*(4), 1305-1314.

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