








Expert Assessment of Synonymic Rows in RuWordNet

Valery Solovyev , Gulnara Gimaletdinova , Liliia Khalitova  ,
and Liliya Usmanova 

Kazan Federal University, Tatarstan 2, 420021 Kazan, Russia
gulnara.gimaletdinova@kpfu.ru, lilia.khalitova@mail.ru

Abstract. This article explores the principles of synsets in the RuWordNet thesaurus and synonyms in the classical dictionaries of Russian synonyms ($N = 10$) to identify discrepancies and improve the principles of organising synsets in RuWordNet. The relevance of the study is determined by the demand for WordNet resources in natural language processing tasks. The authors selected 102 RuWordNet thesaurus synsets, including nouns ($N = 34$), adjectives ($N = 34$) and verbs ($N = 34$). The meanings of the lexemes were correlated according to the data given in Russian language thesauri ($N = 2$). The comparative method and an independent expert assessment of RuWordNet revealed a number of discrepancies and inaccuracies in the representation of synsets concerning polysemy, hypo-hyperonymic relationships, lexical meanings of words and parts-of-speech synonymy. On the basis of this study, the authors recommend the elimination of individual shortcomings in the construction of the RuWord-Net synsets, in particular the polysemy and parts-of-speech synonymy.

Keywords: Computer lexicography · Synonymy · Dictionaries of synonyms · RuWordNet thesaurus · Hypo-hyperonymic relationships

1 Introduction

For the modern stage of linguistic research, the creation of large-scale linguistic resources for information retrieval systems is relevant. The development of such resources is carried out according to a modern approach of linguistic research, computational linguistics, whose development is based on the knowledge of general linguistics. In particular, the development of thesauri takes into account modern advances in lexicology, lexicography, semantics, pragmatics and cognitive linguistics.

WordNet is one of the most popular linguistic resources available today. Developed at Princeton University (<https://wordnet.princeton.edu/>), WordNet is the largest electronic lexical database of English nouns, adjectives, verbs and adverbs. The structure and principles of the organisation of language material, as well as the features of WordNet, are described in detail by a number

of scientific studies [1,2]. In the Google Scholar web search engine, WordNet is mentioned (as of April 30, 2019) in 105,000 articles, which indicates the demand for a resource for scientific research. WordNet is used in various studies in the field of Natural language processing (NLP): information retrieval, automatic text classification, automatic text typing, etc. The lexical database is often used to determine the degree of semantic proximity of words [3] and in the word-sense disambiguation task [4]. Currently, numerous WordNet interfaces, APIs and data processing tools have been developed (<https://wordnet.princeton.edu/related-projects>). WordNet analogues are created for many languages of the world. The Global WordNet Association website <http://globalwordnet.org/> provides information about WordNet-like thesauri for more than 70 languages. A number of studies have described the principles for creating multilingual thesauri [5].

This article is dedicated to the WordNet analogue for the Russian language called RuWordNet [6]. The RuWordNet thesaurus was created at Moscow State University under the guidance of Loukachevitch [7,8]. Currently, it is the only Russian-language thesaurus created by experts and built on the principles of WordNet (synonymic rows and the semantic relationships linking them).

This article presents the results of an independent assessment.

The main contributions of the study are:

1. We provide an independent expert assessment of RuWordNet aimed at the improvement of the quality of RuWordNet synsets.
2. Following the results of the work, we give recommendations to eliminate the discrepancies and inaccuracies revealed in the RuWordNet thesaurus.

The paper is organized as follows. Section 2 provides a brief survey on the related work. Section 3 indicates the data and related methodology. Section 4 provides data analysis and key results. Section 5 discusses the results and gives recommendations. Section 6 concludes the work.

2 Related Work

For the Russian language, several attempts have been made to create thesauri similar to WordNet. The first attempt occurred 20 years ago when the researchers at the philological faculty of St. Petersburg State University launched the RussNet project (<http://project.phil.spbu.ru/RussNet/indexru.shtml>) [9]. RussNet contained 15,000 words and the suggested synsets were described by experts. The project was completed in 2005. According to [10], the project data is not coded uniformly and cannot be used in NLP applications.

The next attempt to create an electronic lexical database was the Russian WordNet resource [11,12] which contained 100,000 words, obtained in a semi-automatic way from various dictionaries. This thesaurus is currently unavailable.

Another project, Wordnet for the Russian language (<http://wordnet.ru/>) was a thesaurus obtained by the automatic translation of WordNet into Russian. This resource which contains 30,000 words is unverified [13].

The Yarn project (<https://russianword.net/>) was a thesaurus created by the crowdsourcing method [10]. Yarn currently contains 145,000 words, but it lacks semantic relationships between synsets, including hypo-hyperonymic, which is typical of thesauri. Both the resource itself and the research based on it are actively developing to date [14–16]. Using Yarn, a selective expert qualitative check of synonymic rows was carried out, where 200 synsets were evaluated by four experts according to the evaluation system, which resulted in 103 synsets of Excellent, 70 of Satisfactory and 27 of Bad quality [10].

The RuWordNet project is developed on the basis of an earlier version called RuThes (<http://www.labinform.ru/pub/ruthes/>). The thesaurus containing 110,000 words and phrases was created by a semi-automatic method on the basis of an extensive corpus of texts with post-editing. While the independent verification of RuWordNet has not been performed, the authors of this paper see this as the purpose of their study. The data on the thesauri is summarized in Table 1.

Table 1. Comparison of WordNet-like thesauri for the Russian language.

Thesaurus	Number of words	Method of creation	Independent verification	Availability	Development stage
RussNet	15,000	Expert-based	No	Partly available	In progress within Yarn project
Russian-WordNet	100,000	Semi-automatic based on dictionaries	No	Unavailable	Completed
Wordnet for the Russian language	30,000	Automatic translation of WordNet into Russian	No	Available	Completed
Yarn	145,000	Crowdsourcing	Yes	Available	In progress
RuWordNet	110,000	Semi-automatic, based on corpus of texts with post-editing	No	Available	In progress

It seems noteworthy that all the thesauri were created by different methods. Accordingly, it is of scientific interest to conduct a comparative analysis of the quality of the created linguistic databases. The analysis of the quality of RuWordNet synsets presented in the article is a step in this direction. The aim of this research is to make an expert assessment of selected RuWordNet synsets with a focus on qualitative analysis.

3 Data and Related Methodology

The current study which presents the analysis of RuWordNet synsets was conducted by independent experts ($N=4$) in Russian semantics and lexicography [17]. First, the experts selected three semantic groups that are supposed to be the most difficult for semantic analysis: a) feelings and emotions, b) mental and verbal activity, and c) human relationships and social life. The raw data consist of 102 synsets from RuWordNet including noun synsets ($N=34$), adjective

synsets ($N = 34$) and verb synsets ($N = 34$). The total number of analysed and compared lexemes is 976 for nouns, 520 for adjectives and 499 for verbs.

Second, the authors chose classical academic dictionaries of Russian synonyms ($N = 10$) [18–27] with different principles for representing synonymic rows, and used a comparative method to analyse the selected synsets in RuWordNet and dictionaries of Russian synonyms particularly considering discrepancies. Thus, a relatively small number of analysed synsets were justified by a qualitative rather than a quantitative approach.

Third, the meanings of the lexemes, mainly those that were polysemantic, were refined in Russian language thesauri ($N = 2$) [28, 29], since numerous cases of discrepancies due to polysemy occurred.

The discrepancies found were summarised and systematised in the form of tables. The full list of words selected for analysis in this study and the tables representing comparative analysis of the synsets are available on the project website (<https://kpfu.ru/kompleksnyj-analiz-struktury-i-soderzhaniya-366287.html>).

Statistical analysis of raw research data allowed the determination of the features of RuWordNet, improved the quality of synsets, as well as the identification and correction of errors in the thesaurus.

4 Results

The RuWordNet thesaurus presents a hierarchical lexeme treatment principle based on hypo-hyperonymic relationships that are established between the generic and species synsets, which are the main structural elements of this thesaurus. One of the basic principles of this resource is the ability to interchange lexical units in most contexts. Moreover, the basic relationships are supplemented by the following: causation and consequence, domain, word formation (single-root words) and parts-of-speech synonymy.

Similarities between the lexemes in RuWordNet synsets and synonymic rows in dictionaries of Russian synonyms were justified if 50% or more dictionaries supported the same meanings, otherwise the lexemes were fixed as discrepancies. This allowed us to make a number of generalisations regarding:

- (1) the description of the polysemy of lexemes;
- (2) a presentation of hypo-hyperonymic relationships;
- (3) the narrowing and extension of the meanings of words;
- (4) a description of parts-of-speech synonymy.

The following are the results of studying the questions above.

4.1 The Polysemy of Lexemes

There are differences in the description of the synsets in RuWordNet and the dictionaries of Russian synonyms. In nine synsets of nouns out of the 34 examined (26%), the lexemes viewed as polysemantic in dictionaries of Russian synonyms

were presented in the RuWordNet thesaurus as monosemantic (*strakh* (fear), *radost'* (joy), *skuka* (boredom), etc.).

A comparative analysis of the synsets in the RuWordNet thesaurus and dictionaries of Russian synonyms revealed significant differences in the description of polysemantic adjectives. Thus, 11 of 34 (32%) RuWordNet synsets are presented as monosemantic, while thesauri [28,29] mark them as polysemantic (adjectives *bezlyudnyy* (deserted), *gostepriimnyy* (hospitable), *neozhidannyy* (unexpected), *truslivyy* (cowardly)).

Similar to nouns, nine synsets out of 34 (26%) demonstrated cases of polysemantic verbs marked as monosemantic in the RuWordNet, which contradicts the descriptions provided by dictionaries of Russian synonym thesauri (the verbs *znat'* (know), *grubit'* (be rude), *mechtat'* (dream), etc.).

4.2 Hypo-hyperonymic Relationships

A comparative analysis of synsets in RuWordNet and dictionaries of Russian synonyms revealed discrepancies in the principles of describing synonymic, hyponymic and hyperonymic relationships. Regarding noun synsets (N = 34), discrepancies in the interpretation of the synonym status between RuWordNet and most dictionaries were noted in 24 synsets (71%): lexemes, defined as synonymic in dictionaries, referred to hyponyms or hyperonyms in the RuWordNet thesaurus. Regarding the total number of analysed nouns (N = 976), 63 cases of discrepancies were identified, which was 6%.

Regarding adjective synsets (N = 34), 23 synsets (68%) lexemes marked as hyponyms or hyperonyms in the RuWordNet thesaurus were viewed as synonyms in most of the analysed dictionaries. Of the total number of analysed adjectives (N = 520), there were 43 such cases (8%).

For verbs (N = 34), similar discrepancies were found in 20 analysed synsets (59%). Regarding the total number of verb lexemes analysed (N = 499), 64 cases of discrepancies per lexeme were identified, which was 11%.

A qualitative analysis of these discrepancies and the interpretation of possible causes are presented in the next section of the article.

4.3 Narrowing and Extension of Meanings

The analysis of the lexemes included in RuWordNet synsets and dictionaries of Russian synonyms revealed some discrepancies in the quantitative and qualitative filling of synonymic rows, which might be explained by different approaches to the interpretation of synonymy and semantic proximity of words in general. We analysed cases of the most significant discrepancies and found the following. First, in the synsets of nouns, adjectives and verbs, there are lexemes which are not represented in the RuWordNet synsets, but are included in the synonymic rows in dictionaries of Russian synonyms. Thus, we can distinguish the narrowing of the lexical meaning when describing lexemes in RuWordNet compared to dictionaries of Russian synonyms. Second, there are also words included in

RuWordNet synsets, that are not presented in the synonymic dictionaries, which we assume to be an extension of the lexical meaning.

The results of a comparative analysis of narrowing and extension of lexical meanings of nouns, adjectives and verbs and statistical data are presented in Fig. 1.

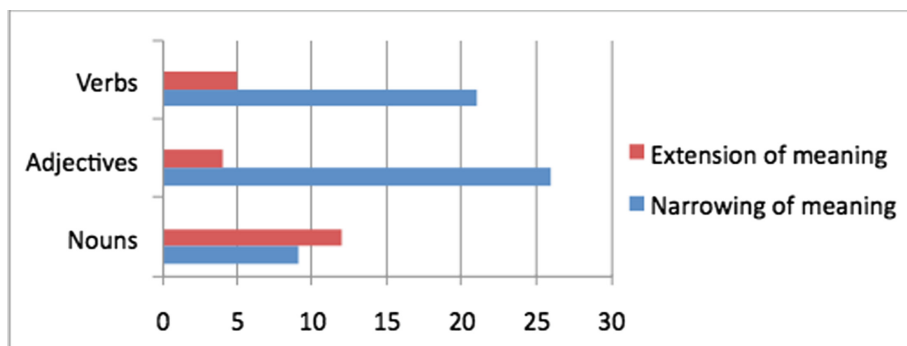


Fig. 1. Narrowing and extension of lexical meanings of nouns, adjectives and verbs.

4.4 Parts-of-Speech Synonymy

As noted earlier, while representing the relationships between words in synsets, RuWordNet suggests the list of part-of-speech synonyms. We were interested in whether parts-of-speech synonymy was presented in all analysed synsets, and whether there were errors or inaccuracies in the description of parts-of-speech synonymy. The analysis revealed the following. Parts-of-speech synonymy in 34 analysed noun synsets was given in 20 cases, which was 59%; however, in the remaining 14 synsets (41%), parts-of-speech synonymy was not included. In the case of adjectives, part-of-speech synonymy was described in 32 analysed synsets (94%) and was not represented in two synsets – only (6%). Of the 34 verb synsets analysed, the parts-of-speech synonyms were given in RuWordNet in 19 cases (56%), while in the remaining 15 verb synsets (44%), parts-of-speech synonyms were not indicated. Inaccuracies in the description of part-of-speech synonymy were found in one noun synset (3%), in four adjective synsets (12%) and in two verbal synsets (6%).

5 Discussion and Recommendations

The question of revising the scope of certain words in RuWordNet remains open, since the problem of hypo-hyperonymic and synonymic relationships between similar lexemes is still debatable and there is no clear answer regarding the semantic status of these units. The analysis revealed significant discrepancies in the description of polysemy: words marked as polysemantic in Russian thesauri

are presented as monosemantic in RuWordNet, which requires clarification and adjustment. For example, nouns *strakh* (fear), *radost'* (joy), *skuka* (boredom), *len'* (laziness), *zhalost'* (pity), *toska* (grief), *mechta* (dream), *obman* (deceit) and *mest'* (revenge).

In RuWordNet the verb *obmanut'* (deceive) is presented with only one meaning: “to deceive, mislead”, while the Russian thesaurus by Ozhegov identifies five meanings of the verb *obmanut'* (deceive): 1. Mislead. 2. Break the promise. 3. Fail to meet expectations/assumptions. 4. Underpay (when calculating wages). 5. Betray, violate marital fidelity [29].

The following meanings are given in the Russian thesaurus by Efremova: 1. Consciously mislead smb. 2. To commit trickery, fraud towards smb. 3. Fail to fulfil your promises, not keep your word. 4. To show deception in love; betray (wife, husband). 5. Seduce (girl, woman) [28].

Recommendations for expanding the meaning of the listed verbs should be considered casual, especially in those controversial cases concerning hyperonymic and synonymic relationships between similar lexemes.

Discrepancies in the principles of describing synonymic, hyponymic and hyperonymic relationships could be illustrated by the following examples (for statistics see Sect. 4).

In the description of the synset *vostorg* (delight) the following words are listed as hyponyms in RuWordNet: *upoyeniye* (flush), *ekstaz* (ecstasy), *ekzal'tatsiya* (exaltation); whereas all the analysed dictionaries of Russian synonyms define them as synonyms [18, 19, 22, 26]. Analysing the synonymic row of the adjective *boyazlivyy* (fearful), we find that seven out of ten dictionaries of Russian synonyms define the relationship between the lexemes *boyazlivyy* (fearful) and *robkiy* (timid) as synonymic; whereas in RuWordNet, *robkiy* is marked as a hyperonym.

We revealed some discrepancies between the words represented in RuWordNet and synonymic rows in the dictionaries of Russian synonyms. For example, the synset for the word *obizhat'* (offend) in RuWordNet contains the following set of synonyms: *zatseplyat'* (hook) and *ushchipyvat'* (pinch), while none of the dictionaries identify them as synonyms. Similarly, the dictionaries do not establish synonymic relationships between *obshchat'sya* (communicate) and *povestis'* (be tricked by); *pridirat'sya* (carp) and *shpynyat'* (poke), *pridirat'sya* (carp) and *podkapyvat'sya* (intrigue); *mechtat'* (dream) and *leleyat'* (cherish); *rasskazyvat'* (tell) and *opisat'* (describe).

Some synsets in RuWordNet in the section “part-of-speech synonymy” have included the words of the same parts of speech as synonyms. For example, the adjectives *belen'kiy* (white) and *veselen'kiy* (cheerful) are given as synonyms of different parts of speech (as nouns) to the adjectives *belyy* (white) and *veselyy* (funny), respectively. We assume that this is due to the phenomenon of substantivisation; however, we recommend clarifying the part-of-speech synonyms of these words. Regarding verbs, the words *znat'* (know) and *nadsmekhat'sya* (make fun of) require clarification.

The cases of narrowing of the meanings of synonyms in RuWordNet compared to the dictionaries of Russian synonyms are explained since RuWordNet was

based on a news corpus, while the classical dictionaries of synonyms were focused on fiction. However, we believe that, with the expansion of the corpus and the inclusion of fiction, the list of synsets in RuWordNet will increase, while the discrepancies will decrease.

The cases of expansion of the meanings of synonyms in RuWordNet are likely related to a larger number of words (110,000) compared to the dictionaries of Russian synonyms, covering a significantly smaller layer of vocabulary.

Evaluating the quality of hypo-hyperonymic relationships in RuWordNet is an extremely complicated task, primarily due to the lack of a formal (operational) definition of hypo- and hyperonymy in linguistics. Modern computational linguistics also does not provide methods for the automatic detection of hypo-hyperonymic relationships corresponding to “golden standards”. Moreover, there is no elaborate Russian language thesaurus compiled by professional lexicographers. This could be beneficial in comparing and analysing ambiguous data. To establish valid hypo-hyperonymic relationships, we recommend carrying out extensive theoretical studies that go far beyond the scope of this article.

6 Conclusions

In this research, we analysed the synsets presented in RuWordNet thesaurus and compared the data with dictionaries of synonyms of the Russian language. The comparative method and an independent expert assessment of RuWordNet revealed a number of discrepancies and inaccuracies in the representation of synsets concerning polysemy, hypo-hyperonymic relationships, lexical meanings of words and parts-of-speech synonymy. The recommendations would be beneficial in the creation and improvement of similar linguistic databases, and expert assessment seems to be the most appropriate approach in cases when qualitative analysis is needed.

Acknowledgments. This research was financially supported by the Russian Foundation for Basic Research (Grant No. 18-00-01238), and by the Government Program of Competitive Development of Kazan Federal University.

References

1. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**(11), 39–41 (1995)
2. Fellbaum, C.: WordNet and wordnets. In: Brown, K., et al. (eds.) *Encyclopedia of Language and Linguistics*, 2nd edn, pp. 665–670. Elsevier, Oxford (2005)
3. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity – measuring the relatedness of concepts. In: *Demonstration Papers at HLT-NAACL 2004*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 38–41 (2004)
4. Navigli, R.: Word sense disambiguation: a survey. *ACM Comput. Surv.* **41**(2), 1–69 (2009)
5. Vossen, P.: EuroWordNet: building a multilingual database with wordnets for European languages. *The ELRA Newsl.* **3**(1), 7–10 (1998)

6. Thesaurus of Russian Language RuWordNet. <https://ruwordnet.ru/ru>. Accessed 26 May 2019
7. Loukachevitch, N.: Thesauri in Information Retrieval Tasks. Moscow University, Moscow (2011)
8. Loukachevitch, N., Lashevich, G., Gerasimova, A., Ivanov, V., Dobrov, B.: Creating Russian WordNet by conversion. In: Proceeding of Conference on Computational Linguistics and Intellectual Technologies Dialogue-2016, RSUH, Moscow, pp. 405–415 (2016)
9. Azarova, I.V., Sinopalnikova, A.A., Yavorskaya, M.V.: Guidelines for RussNet structuring. In: Proceeding of Conference on Computational Linguistics and Intellectual Technologies Dialogue-2004, Nauka, Moscow, pp. 542–547 (2004)
10. Braslavski, P., Ustalov, D., Mukhin, M., Kiselev, Y.: YARN: spinning-in-progress. In: Proceedings of the 8th Global WordNet Conference, Bucharest, Romania, pp. 58–65 (2016)
11. Balkova, V., Sukhonogov, A., Yablonsky, S.: Russian WordNet from UML-notation to internet/intranet database implementation. In: Proceedings of the 2nd International WordNet Conference, Masaryk University, Brno, pp. 31–38 (2004)
12. Sukhonogov, A., Yablonsky, S.: Russian WordNet development. In: Proceedings of the 6th Russian Conference on Digital Libraries: Advanced Methods and Technologies, Digital Collections - RCDL 2004, Pushchino, Russia (2004)
13. Gelfenbeyn, I., Goncharuk, A., Lekhelt, V., et al.: Automatic translation of WordNet semantic network to Russian language. In: Proceeding of Conference on Computational Linguistics and Intellectual Technologies Dialogue-2003 (2003)
14. Ustalov, D., Chernoskutov, M., Biemann, C., Panchenko, A.: Fighting with the sparsity of synonymy dictionaries for automatic synset induction. In: van der Aalst, W.M.P., Ignatov, D.I., Khachay, M., Kuznetsov, S.O., Lempitsky, V., Lomazova, I.A., Loukachevitch, N., Napoli, A., Panchenko, A., Pardalos, P.M., Savchenko, A.V., Wasserman, S. (eds.) AIST 2017. LNCS, vol. 10716, pp. 94–105. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73013-4_9
15. Ustalov, D., Teslenko, D., Panchenko, A., Chernoskutov, M.: Mnozoznal: an unsupervised system for word sense disambiguation. In: Proceedings of 2017 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), pp. 147–150. IEEE, Novosibirsk (2017)
16. Ustalov, D.: Expanding hierarchical contexts for constructing a semantic word network. In: Proceeding of Conference on Computational Linguistics and Intellectual Technologies Dialogue-2017, RSUH, Moscow, pp. 369–381 (2017)
17. Janda, L., Solovyev, V.: What constructional profiles reveal about synonymy: a case study of Russian words for sadness and happiness. *Cogn. Linguist.* **20**(2), 367–393 (2009)
18. Abramov, N. (ed.): Dictionary of Russian Synonyms and Similar Expressions, 7th edn. Russkiye Slovare, Moscow (1999). (in Russian)
19. Aleksandrova, Z. (ed.): Dictionary of Synonyms of the Russian Language: A Practical Guide, 15th edn. Russkiy yazyk, Moscow (2007). (in Russian)
20. Alektorova, L., Vvedenskaya, L., Zimin, V., et al. (eds.): Dictionary of Synonyms of the Russian Language, 2nd edn. Astrel, AST, Moscow (2002). (in Russian)
21. Gorbachevich, K. (ed.): A Brief Dictionary of Synonyms of the Russian Language. Eksmo, Moscow (2005). (in Russian)
22. Kozhevnikov, A. (ed.): Large Dictionary of Synonyms of the Russian Language: Speech Equivalents: A Practical Guide. Neva, St. Petersburg (2003). (in Russian)
23. Klyueva, V. (ed.): A Brief Dictionary of Synonyms of the Russian Language. Uchpedgiz, Moscow (1961). (in Russian)

24. Apresyan, Y. (ed.): A New Explanatory Dictionary of Synonyms of the Russian Language, 2nd edn. Yazyki russkoy kultury, Moscow (2000). (in Russian)
25. Babenko, L. (ed.): Dictionary of Synonyms of the Russian Language. Astrel, Moscow (2011). (in Russian)
26. Evgenyeva, A. (ed.): Dictionary of Synonyms: Reference Book. Nauka, Leningrad (1975). (in Russian)
27. Dictionary of Synonyms of the Russian Language: Dictionary of Antonyms of the Russian Language. Victoria Plus, St. Petersburg (2007). (in Russian)
28. Efremova, T.: The New Dictionary of the Russian Language. Thesaurus and Word-Building. Russkiy yazyk, Moscow (2000). (in Russian)
29. Ozhegov, S.: Russian thesaurus. <https://slovarozhegova.ru/>. Accessed 26 April 2019 (in Russian)