# IDENTIFYING MORPHOLOGICAL PROPERTIES OF RUSSIAN WORDS WITH THE ONTOLOGY-BASED TEXT ANALYSER

**Ksenia Balysheva[1]*, Elena Kartashova[2], Konstantin Kondratiev[3], Aleksey Mikheev[4]**

[1]Asst. Prof. Dr., Mari State University, Russia, qsuaka@mail.ru
[2] Prof. Dr., Mari State University, Russia, elena.karta77@mail.ru
[3] Project Manager, Telephone Systems Ltd, Russia, kk@digt.ru
[4] Ph. D. Student, Mari State University, Russia, scurra.42@yandex.ru
*Corresponding Author

## Abstract

This article presents the first stage of an ongoing effort of creating the application Ontology-Based Text Analyser (OTA) aimed at automatically identifying semantics and grammatical properties of widely used Russian words in connected texts. At present this application identifies only morphological properties of Russian words. In this application all morphological properties of content words and grammatical function words are revealed on the basis of a query to the Ontology of Russian Grammatical Forms (OntoRuGrammaForm) that we set up earlier. In OntoRuGrammaForm we used LexInfo[1] which represents morphological properties of words in the ontological format as a scheme for data organising. To set up OntoRuGrammaForm the existing LexInfo ontology was extended with missing and refined grammatical categories. In OntoRuGrammaForm the linkage of semantics with morphological properties is implemented with OntoLex[2] which makes it possible to link grammatical word forms with lemmas and lemmas with concepts in knowledge area ontologies. The automatic process of word morphology identification is illustrated with a connected text of the informative type taken from the open news online-portal. In this news text the system of morphological properties of words is identified with OntoRuGrammaForm. This application also displays lemmas and transcription of separate words in a connected text. The created application (OTA) can be used as an innovative methodical tool in teaching Russian to foreign students to develop skills of identifying morphological characteristics of words in texts. At present this application is available on the Web in open access and can be used for analysing morphological properties of widely used Russian words in a connected text.

**Keywords**: Automatic identifying, Morphological properties, Ontology, Ontology-based application, Text analyser, LexInfo, OntoLex

[1] https://www.lexinfo.net
[2] https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

## 1. INTRODUCTION

In computational linguistics one of the approaches to the representation of natural language properties is the ontological approach. This approach is currently being developed now, mainly in researching natural language processing. Despite this, there is no unanimity in defining the term "ontology" among researchers (see Gruber, 1993, Guarino, 1998, Guarino and Welty, 2004, Loukachevitch, 2010, Studer et al., 1998). In this article by the term "ontology" we mean a formal explicit description of concepts in a machine-readable and interpretable format.

The topicality of the ontological approach is based on the fact that on the Semantic Web there exist ontological models representing linguistic Linked Data that describe morphological features of languages to some extent, including Russian, e.g., *OliA* (Chiarcos, 2008), *lemon* (McCrae et al., 2011), *LexInfo* (Cimiano et al., 2011). Representation of features of a natural language as ontologies on the Semantic Web makes it easier to implement the idea of the Linked Data, which has led to the emergence of the Linguistic Linked Open Data (LLOD) cloud[3], a cross-domain knowledge base comprising structured information extracted from Wikipedia infoboxes, the World Atlas of Language Structures (WALS)[4] and lexical resources such as Wiktionary[5], WordNet, FrameNet (Cimiano et al., 2014) and BabelNet (Navigli et al., 2010). The advantages of the Linked Data for linguistics include representational adequacy, structural and conceptual interoperability, data federation (see Chiarcos et al., 2013).

Another advantage of representing data in an ontology-based form is its contribution to creating and expanding conversational interface capabilities. Ontologies enable setting up a system of a database query to a certain knowledge domain in a natural language.

The Ontology-Based Text Analyser (OTA) is the application created for the initial processing of connected texts in Russian. Using this application, one can identify morphological properties of Russian content words (nouns, adjectives, verbs, adverbs, pronouns, participles, and gerunds) and grammatical function words (prepositions, conjunctions, and particles). According to the separability hypothesis (Miller, 1998), we can study and describe morphological and syntactical properties of a language apart from its lexical and semantic levels. The identification of morphological properties of words is essential for further ascertaining semantic and syntactical relations in phrases and sentences of connected texts in a natural language in general and the Russian language in particular.

In this application (OTA) the morphological properties of Russian content words and grammatical function words are identified with the Ontology-Based Dictionary of Russian Grammatical Forms (OntoRuGrammaForm) that we created earlier. For labelling OntoRuGrammaForm we used LexInfo while for relating concepts in a dictionary entry OntoLex was used.

In this article the work of the OTA application is illustrated with a connected text of the informative type under the headline "The Pushkin Museum takes part in a biannual exhibition in Venice for the first time" taken from the Russian Ria Novosti news portal[6]. All the texts that have been analysed with OTA were taken from Russian news portals as in news texts words are basically used in their direct meanings and the number of stylistic and artistic means is limited.

## 2. CREATING THE ONTOLOGY-BASED DICTIONARY OF RUSSIAN GRAMMATICAL FORMS

The idea of connecting words with concepts, including the morpho-syntactic level, makes it possible to clarify the meaning, e.g., of polysemantic and homonymous words. This idea is implemented in LexInfo. LexInfo, the part of OntoLex, is a universal multipurpose model for representing morpho-syntactic properties of highly inflected languages, including Russian. The accomplished analysis of the LexInfo structure showed that the morpho-syntactic properties of Russian are not fully represented in LexInfo. So, the first step of creating the Ontology-Based Dictionary of Russian Grammatical Forms was adjusting the morpho-syntactic properties, given in LexInfo, in accordance with the state-of-the-art grammar of the Russian literary language. With the additions and adjustments, introduced into LexInfo, it became possible to represent morpho-syntactic properties of Russian more completely and accurately in the Ontology-Based Dictionary of Russian

[3] https://linguistics.okfn.org/llod
[4] https://wals.info
[5] https://en.wiktionary.org/wiki
[6] https://ria.ru/culture/20170503/1493568991.html

Grammatical Forms (OntoRuGrammaForm).

As a source for OntoRuGrammaForm we used the Open Corpora[7], the open corpus of the Russian language. The Open Corpora is compiled by volunteers using web texts and is available in XML and plaintext formats. The Open Corpora XML schema can be viewed at https://opencorpora.org/export/dict/dict.opcorpora.xsd. The automatic conversion of the Open Corpora labels into the OntoLex labels is a 1:1 mapping. The project of label conversion is available at https://github.com/cnstntn-kndrtv/opencorpora2ontolex.

In the Ontology-Based Dictionary of Russian Grammatical Forms (OntoRuGrammaForm) the data is serialized in HDT. Like RDF/XML, HDT is a format for RDF, but it keeps datasets compressed. OntoRuGrammaForm is available for public use as the Linked Data Fragments[8] access point at http://ldf.kloud.one/ontorugrammaform.

At present OntoRuGrammaForm contains 389,360 entries and 5,489,044 word forms. Each dictionary entry is related to a lemma and word forms. The morpho-syntactic properties of words are described with the LexInfo model. As an example we give Russian lexemes: 1) ‘мир’ (‘mir’), similar to the English lexeme ‘world’, which means “the totality of all substance types”; 2) ‘миро’ (‘miro’), similar to the English lexeme ‘chrism’, which means “a mixture of oil and balsam, consecrated and used for anointing at baptism and in other rites in Christianity” (Ozhegov, 1983). These are two different lexemes, each having its own set of morphological forms. These two words have homonymous word forms, one of them – “mira” – is used in the exemplifying text showing the principle of operation of the developed application. Fig. 1 below shows the description of the word ‘мир’ (‘mir’) – ‘world’, its lemma and three forms.

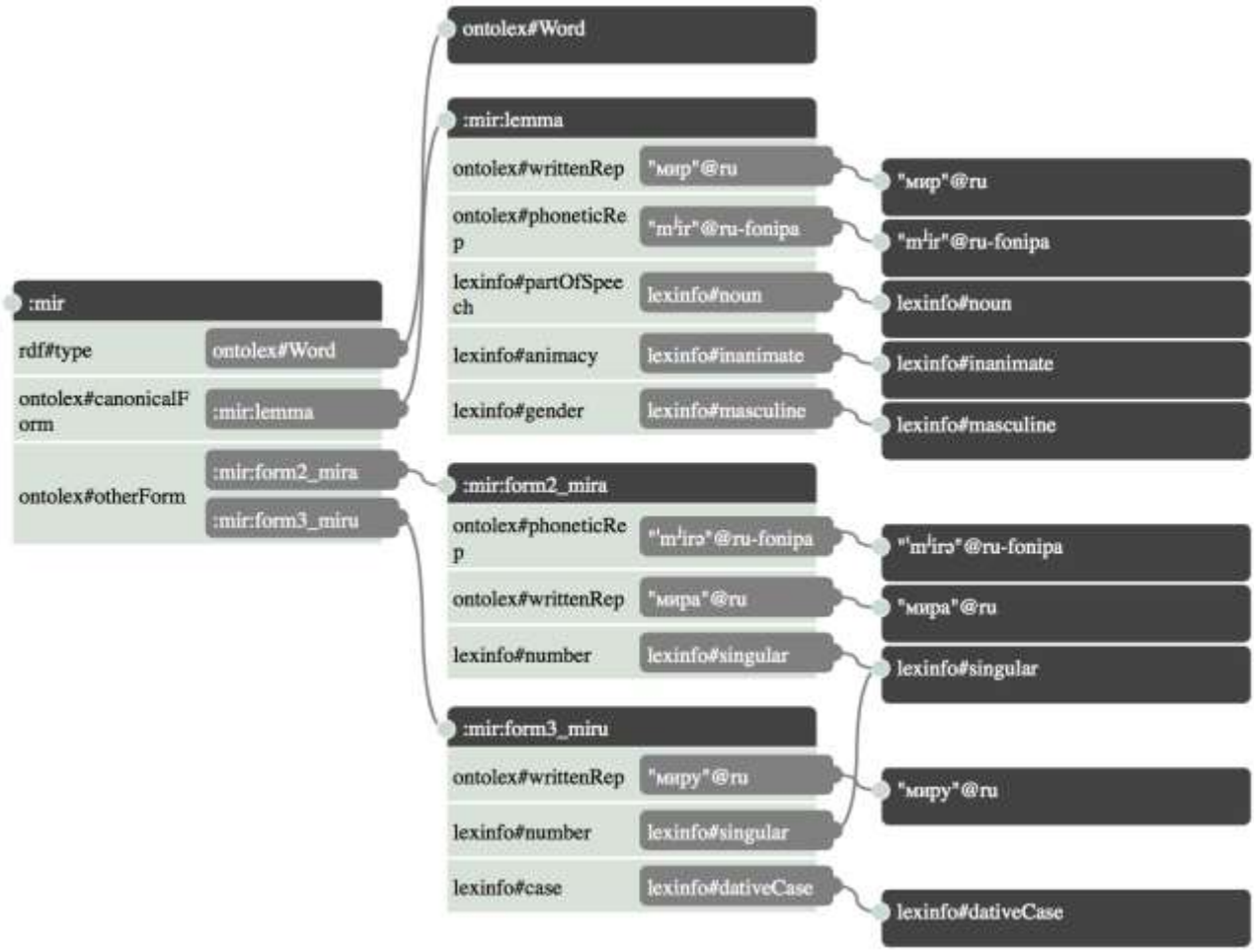



Fig. 1. Visualisation of relations between the morphological forms of the word ‘мир’ (‘mir’) – ‘world’

[7] https://opencorpora.org
[8] https://linkeddatafragments.org

## 3. DESCRIPTION OF THE OTA APPLICATION DESIGN AND PRINCIPLE OF OPERATION

The Ontology-Based Application (OTA) is available at https://sw.kloud.one. This application consists of two blocks: the dictionary and the user application. The description of the dictionary (OntoRuGrammaForm) is given in the previous section.

The backend part of the user application is written in Node JS[9] and the Express[10] framework employing the HTML template engine PUG[11]. In our view, this architectural solution will enable the functionality of the application to be developed regularly in the future. The user part of the application is designed with a possibility to expand its functionality and consists of three main modules: 1) database query module; 2) text editor; 3) query result output module. The application user interface is shown in Fig. 2.

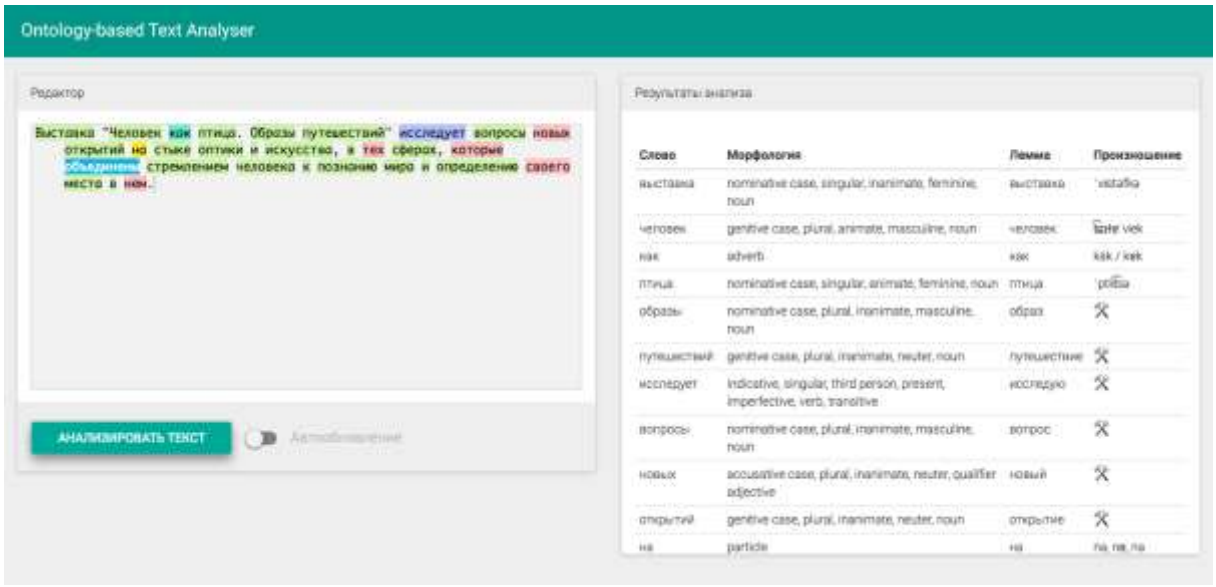



Fig. 2. Application user interface

### 3.1. Database query module

The database query module is based on the Linked Data Fragments Client[12] and is the JavaScript code, which is started up simultaneously with the Web Workers technology. This fact makes it possible to fulfil a few queries at the same time (the number of workers equals the number of cores of a user's CPU).

### 3.2. Text editor

The text editor is based on the ACE editor[13] but with several modifications. With these modifications it is feasible to run changes of syntax highlighter rules dynamically in the editor. Each part of speech is highlighted with its own colour. When a user looks at a highlighted text it is evident at a glance what part of speech dominates in a processed text. In the exemplifying text the majority of words are nouns (Fig.3).

[9] https://nodejs.org
[10] https://expressjs.com
[11] https://pugjs.org
[12] https://github.com/LinkedDataFragments/Client.js
[13] https://ace.c9.io

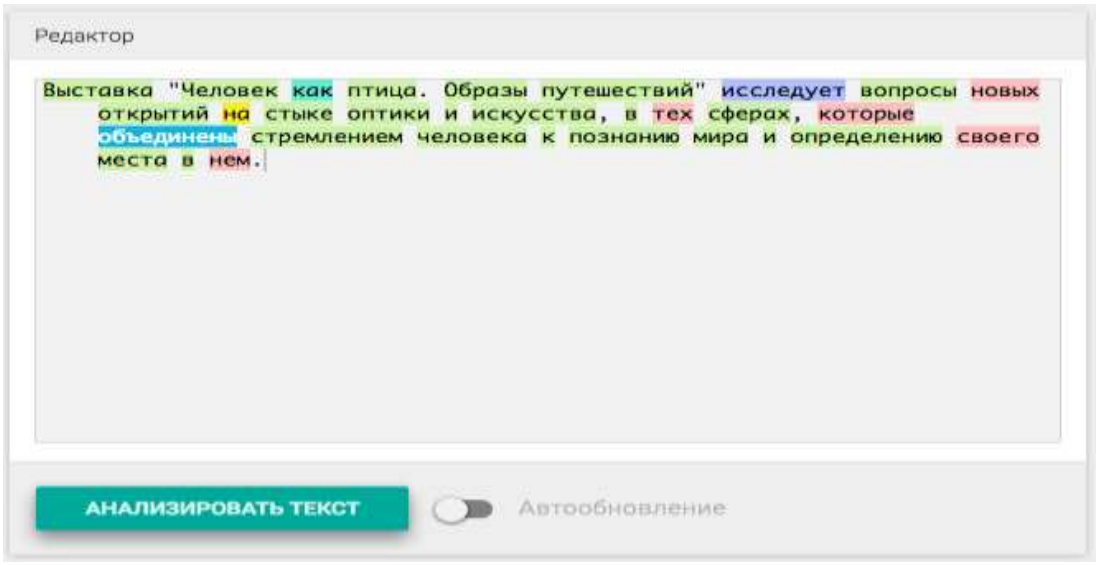
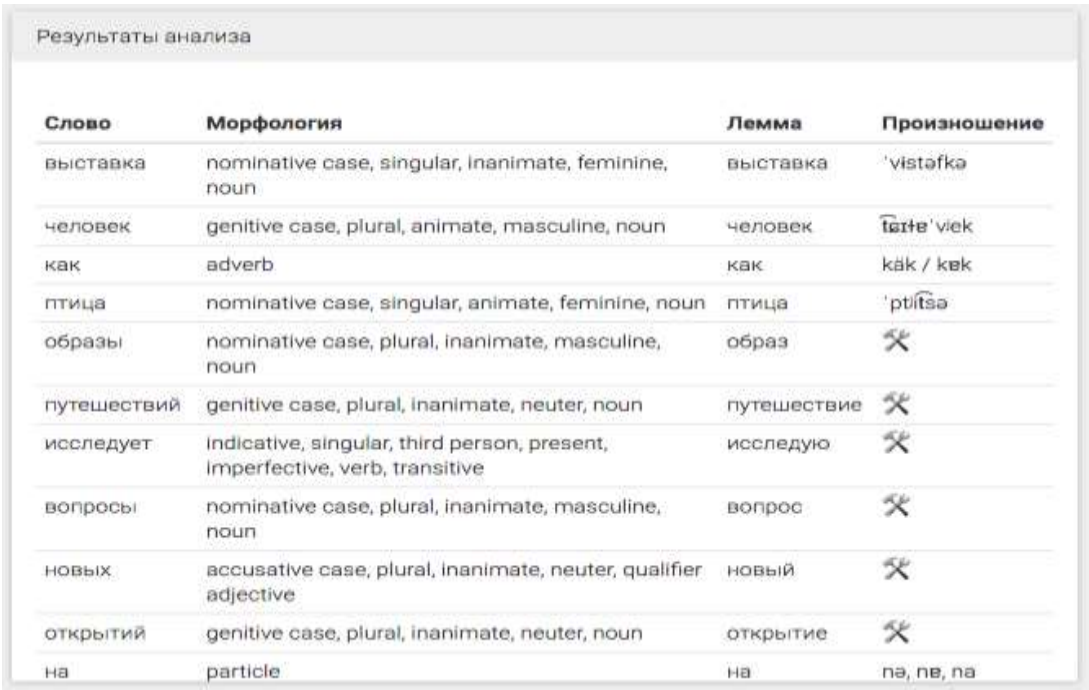



Fig. 3. Text editor window

## 3.3. Query result output module

The query result output module is represented by the table, which cells are updated as results arrive from the database query module. The table is based on the D3.js library[14]. The table contains the following columns (Fig. 4): 1) “word” which is a word form as it is given in the analysed text; 2) “morphology”, i.e., morphological properties of a word form; 3) “lemma”, i.e., a root form; 4) “pronunciation”. A word form and a lemma (columns 1 and 3) are given in Russian. A list of morphological properties of words (column 2) is given in English for the convenience of foreign learners of Russian. Pronunciation of a word (column 4) comes in the IPA[15] phonetic transcription. At present only about 10% of the analysed word forms are supplied with transcription but this list is being developed.


Результаты анализа

| Слово | Морфология | Лемма | Произношение |
|---|---|---|---|
| выставка | nominative case, singular, inanimate, feminine, noun | выставка | ˈvistəfkə |
| человек | genitive case, plural, animate, masculine, noun | человек | t͡ɕɪɪ̯ɪˈvʲek |
| как | adverb | как | käk / kʁk |
| птица | nominative case, singular, animate, feminine, noun | птица | ˈpt͡ɕɪt͡sə |
| образы | nominative case, plural, inanimate, masculine, noun | образ | 🛠 |
| путешествий | genitive case, plural, inanimate, neuter, noun | путешествие | 🛠 |
| исследует | indicative, singular, third person, present, imperfective, verb, transitive | исследую | 🛠 |
| вопросы | nominative case, plural, inanimate, masculine, noun | вопрос | 🛠 |
| новых | accusative case, plural, inanimate, neuter, qualifier adjective | новый | 🛠 |
| открытий | genitive case, plural, inanimate, neuter, noun | открытие | 🛠 |
| на | particle | на | nə, nʁ, na |

Fig. 4. Analysis result window

[14] https://d3js.org
[15] International Phonetic Alphabet

## 4. DISCUSSION AND RESULTS

The testing of the application was accomplished on 100 texts of the informative type taken from Russian news portals at random. According to the procedure algorithm of the application, it automatically chooses the first word form with its morphological properties from the list of homonymous word forms. The upgrading of the current application will involve semantics and syntax analysis. With the syntactical structure of phrases and sentences taken into account analysis results with the application will be considerably more accurate.

While testing the application with 100 news texts, no such case occurred that a word form was not found on a query into OntoRuGrammaForm. This high effectiveness of the application operation is determined by the inclusive and updating character of OntoRuGrammaForm. As it is based on the OpenCorpora data, it includes literary words as well as slang and jargon lexemes, dialect lexemes, and other substandard lexemes. The number of entries is now 389,360 and it is constantly growing. To compare, the Great Academic Dictionary of the Russian language counts about 150,000 entries (Kruglikova, 2012). The researchers estimate that approximately 400,000 words exist in the Russian language at the present stage of its development[16]. Thus, the current version of OTA with the query into OntoRuGrammaForm covers the majority of existing Russian words.

The OTA application is a convenient methodical tool which can be used both in class and for independent work in teaching Russian to foreign students. The two evident advantages of this application for foreign learners that enable it to be used for self-control and self-evaluation include the following: 1) morphological properties of word forms appear in English; 2) the transcription of word forms is given in the IPA symbols.

## 5. CONCLUSION AND FUTURE WORK

The Ontology-based Text Analyser is the application that is currently capable of identifying morphological properties of words in a connected text. It is being extended into the application aimed at identifying semantics and syntax of a connected text. The accuracy of morphological property identification without considering syntactical structure of phrases and sentences is quite sufficient. When a syntactical structure of sentences will be taken into account the accuracy of analysis will considerably improve. The options that are being developed now include the following: 1) meanings and phonetic transcription of words; 2) semantic and syntactical analysis of whole sentences and texts.

This application was tested on news texts. The application accuracy of identifying semantics, morphology and syntax of word forms in a connected text will also be tested on texts of other functional styles and genres.

## 6. ACKNOWLEDGEMENT

The authors are grateful to Telephone Systems Ltd for support and technical assistance as a part of kloud.one project.

## REFERENCE LIST

Chiarcos, C. (2008). An ontology of linguistic annotations. LDV Forum, pp. 1–136.

Chiarcos, Ch., McCrae, J., Cimiano, Ph., Fellbaum, Ch. (2013). Towards open data for linguistics: linguistic linked data, new trends of research in ontologies and lexical resources. Springer.

Cimiano, P., McCrae, J., Buitelaar, P., Stintek, M. (2011). Lexinfo: A declarative model for the lexicon-ontology interface. Web Semantics: Science, Services and Agents on the World Wide Web, pp. 29–51.

Cimiano, Ph., Unger, Ch., McCrae, J. (2014). Ontology-based interpretation of natural language.

Gruber, T. (1993). Toward principles for the design of ontologies used for knowledge sharing. Formal Analysis in Conceptual Analysis and Knowledge Representation. Kluwer.

Guarino, N. (1998). Some ontological principles for designing upper level lexical resources. Proceedings of First International Conference on Language Resources and Evaluation (LREC), pp. 527–534.

[16] https://rg.ru/2014/10/10/slovari.html

Guarino, N., Welty, Ch. (2004). An overview of OntoClean, handbook on ontologies. Springer., pp. 151–159, Springer.

Kruglikova, L.E. (2012). The Great academic dictionary of the Russian language as a successor of Russian academic lexicography traditions. Notebooks of Spanish Studies of Russian Philology, vol. 8, pp. 177 – 198.

Loukachevitch, N.V. (2010). Thesauri in tasks of information retrieval. Moscow.

McCrae, J., Spohr, D., and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. The Semantic Web: Research and Applications, pp. 245–259

Miller, G. (1998). Nouns in WordNet, Fellbaum, C (ed) WordNet, An Electronic Lexical Database, The MIT Press, pp.23-47.

Navigli, R., Ponzetto, S. (2010) BabelNet: building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 216 – 225.

Ozhegov, S.I. (1983). Dictionary of the Russian language (in Russian). Moscow.

Studer, R., Benjamins, V., and Fensel, D. (1998). Knowledge engineering: Principles and methods. Data Knowledge Engineering, pp. 161–197.