Abstract. This article describes possibilities of using OntoLex as a model for creating an ontology of morpho-syntactic properties of the Russian language. For this purpose we analysed morpho-syntactic properties of Russian, given in LexInfo and then extended it with grammatical categories that are not represented or that are not correctly defined in LexInfo. The introduced supplements and adjustments enable LexInfo to represent morpho-syntactic properties of the Russian language more completely and to use it for creating the Ontology-Based Dictionary of Russian Grammatical Forms (OntoRuGrammaForm). The created ontology-based dictionary helps to detect grammatical forms of widely used Russian words.

Keywords: OntoLex, LexInfo, Ontology, Morpho-syntactic properties, Ontology-Based Dictionary of Russian Grammatical Forms (OntoRuGrammaForm).

1 Introduction

The ontological approach to representation of natural language properties is currently being developed in computational linguistics, mainly in researching natural language processing. On the Semantic Web there are various ontology-based lexical and semantic datasets, e.g. WordNet [8], FrameNet [2], BabelNet [12], RussNet [1], RuThes [9], RuWordNet [10], YARN [3].

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 [4], lemon [11], LexInfo [6]. 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\(^1\), a cross-domain knowledge base comprising structured information extracted from Wikipedia infoboxes, the World Atlas of Language Structures (WALS)\(^2\) and lexical resources such as Wiktionary\(^3\), WordNet, FrameNet [7] and BabelNet. The advantages of the Linked Data for linguistics include representational adequacy, structural and conceptual interoperability, data federation [5].

The idea of connecting words with concepts, including the morpho-syntactic level, which makes it possible to clarify the meaning, e.g. of polysemic and homonymous words, is implemented in LexInfo. In this project we used LexInfo as the most complete ontology based on RDF model for labeling the Ontology-Based Dictionary of Russian Grammatical Forms due to its evident advantages: separation and independence between the ontological and linguistic levels; structuring linguistic information; the ability to specify the meaning of linguistic constructions with respect to arbitrary ontologies, etc. [6]. In LexInfo the data is serialized in RDF/XML, while in OntoRuGrammaForm the data is serialized in HDT. Like RDF/XML, HDT is a format for RDF, but it keeps datasets compressed.

The goal of this project is to create an ontology-based dictionary that represents morpho-syntactic properties of the Russian language. To achieve this goal we set and consecutively resolved the following tasks: 1) analysing grammatical classes and properties of Russian, given in LexInfo; 2) collating the composition of grammatical classes and properties in LexInfo with Russian grammar books and dictionaries; 3) supplementing LexInfo with insufficient and refined Russian grammatical categories; 4) translating labels into Russian and supplying LexInfo and OntoLex elements with Russian commentaries; 5) creating the Ontology-Based Dictionary of Russian Grammatical Forms.

Both LexInfo and OntoLex were used to create the Ontology-Based Dictionary of Russian Grammatical Forms. Grammatical categories of words were determined with LexInfo, while entities/concepts in a dictionary entry were related with OntoLex.

2 Supplementing the LexInfo Model with Russian Grammatical Categories

LexInfo is a universal multipurpose model for representing morpho-syntactic properties of highly inflected languages that have genetic and typological resemblances at the level of common affixes, roots, and a regular phonetic correspondence of sounds. In general, morpho-syntactic properties of Russian can be represented in LexInfo. Nevertheless, the accomplished analysis of its structure showed that these properties are not fully represented. This fact gave rise to the intent of adjusting these properties, listed in LexInfo, in accordance with the state-of-the-art of grammar of the Russian literary language.

\(^1\) http://linguistics.okfn.org/lloid
\(^2\) http://wals.info
\(^3\) https://en.wiktionary.org/wiki
The analysis of the list of Russian grammatical properties in LexInfo and its collation with the data of academic grammar books [14, 15] led to the following observations:

1) some grammatical categories of Russian are not represented and do not have special nominations in LexInfo;
2) some grammatical categories are not placed into correct grammatical classes/properties;
3) some grammatical categories are supplied with inaccurate Russian translations.

The analysis of LexInfo showed that nominations of some Russian grammatical categories should be introduced (see Table 1):

1) In LexInfo the individual *participle* is put into the class *VerbFormMood*. In our view, it should also belong to the class *PartOfSpeech*. So, we introduced the new class *ParticiplePOS*, into which the individual *participle* is placed.
2) To the class *ParticiplePOS* we added the new individual *shortParticiple*. The distinction between a short participle and a participle is essential for the system of the Russian language as these two forms have different inflections and different syntactical functions.
3) In LexInfo there is no individual *gerund*. We believe it should be added to identify the adverbial participle (the Russian gerund) as the part of speech in Russian. We introduced the new class *GerundPOS*, into which the individual *gerund* is put, and we also stated that the individual *gerund* belongs to the class *VerbFormMood*.
4) We added the individuals singulariaTantum, pluraliaTantum, fixedNumber to the existing class Number.
5) We added the new class Finiteness with two individuals – finite and nonFinite – to the class MorphosyntacticProperty.
6) We introduced the class Reflexivity with two individuals – reflexive and nonReflexive into the class MorphosyntacticProperty.
7) The individual *impersonalVerb* is added to the class *VerbPOS*.
8) The individual *shortAdjective* is added to the class *AdjectivePOS*.
9) The individual *relativeAdjective* is added to the class *AdjectivePOS*.
10) The individual *collectiveNumeral* is added to the class *NumeralPOS*.

The supplementation of grammatical categories of the Russian language in LexInfo is also connected with eliminating inaccuracies in placing grammatical categories into classes (see Table 1):

1) In LexInfo *comparative* is the individual of the class *Degree*. In our view, it is also the individual of the class *AdjectivePOS*.
2) In LexInfo the individual *infinitive* belongs to the class *VerbFormMood*. In our view, it also belongs to the class *VerbPOS*.
3) In LexInfo the individual *ordinalAdjective* belongs to the class *AdjectivePOS*. According to the grammatical properties of Russian this individual also belongs to the class *NumeralPOS*.
Another important supplement to grammatical properties of Russian in LexInfo is adjusting translations of class and individual labels into Russian. Some examples of this type of supplements are given below:

(1) The term gerundive, which is put into the class VerbFormMood, is not accurately translated into Russian. In Latin the gerundive is a verbal adjective while the gerund is a verbal noun both in Latin and in English. In Russian the grammatical category of a gerund does not exist. We suggested introducing the individual gerundPOS to label the adverbial participle (the Russian gerund) as the part of speech.

(2) In Russian there exist cardinal numerals and ordinal numerals. In LexInfo the Russian labels for the individuals cardinalNumeral and ordinalNumeral from the class NumericalPOS are confused and should be interchanged.

(3) In LexInfo class Finiteness from the class MorphosyntacticProperty is labeled inaccurately in Russian. Our suggestion is to supply the grammatical category of finiteness as well as the class Finiteness by the Russian label spryagaemost. As the English conjugation and the Russian spryagaemost are quasi-synonyms, we find the LexInfo label Finiteness appropriate to indicate the ability of Russian verbs to conjugate.

Table 1. Suggested supplements to LexInfo for representing grammatical categories of Russian.

<table>
<thead>
<tr>
<th>№</th>
<th>Individual</th>
<th>Class</th>
<th>Commentary on supplements</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>participle</td>
<td>VerbFormMood &amp; ParticiplePOS</td>
<td>The individual participle belongs to the class verbFormMood. The new class ParticiplePOS is added. The individual participle should belong to the class ParticiplePOS and to the class VerbFormMood.</td>
</tr>
<tr>
<td>2</td>
<td>shortParticiple</td>
<td>VerbFormMood &amp; ParticiplePOS</td>
<td>The new individual shortParticiple is added to the class ParticiplePOS. It should belong to both classes - VerbFormMood and ParticiplePOS.</td>
</tr>
<tr>
<td>3</td>
<td>gerund</td>
<td>VerbFormMood &amp; GerundPOS</td>
<td>The new individual gerund is added to two existing classes – VerbFormMood and GerundPOS.</td>
</tr>
<tr>
<td>4</td>
<td>singulariaTantum</td>
<td>Number</td>
<td>The new individual singulariaTantum is added to the existing class Number.</td>
</tr>
<tr>
<td>5</td>
<td>pluraliaTantum</td>
<td>Number</td>
<td>The new individual pluraliaTantum is added to the existing class</td>
</tr>
</tbody>
</table>
Number.

<table>
<thead>
<tr>
<th></th>
<th>fixedNumber</th>
<th>Number</th>
<th>The new individual fixedNumber is added to the existing class Number.</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>finite</td>
<td>Finiteness</td>
<td>The new individual finite and the class Finiteness are added.</td>
</tr>
<tr>
<td>8</td>
<td>nonFinite</td>
<td>Finiteness</td>
<td>The new individual nonFinite and the class Finiteness are added.</td>
</tr>
<tr>
<td>9</td>
<td>reflexive</td>
<td>Reflexivity</td>
<td>The new individual reflexive and the class Reflexivity are added.</td>
</tr>
<tr>
<td>10</td>
<td>nonReflexive</td>
<td>Reflexivity</td>
<td>The new individual nonReflexive and the class Reflexivity are added.</td>
</tr>
<tr>
<td>11</td>
<td>impersonalVerb</td>
<td>VerbPOS</td>
<td>The new individual impersonalVerb is added to the existing class VerbPOS.</td>
</tr>
<tr>
<td>12</td>
<td>shortAdjective</td>
<td>AdjectivePOS</td>
<td>The new individual shortAdjective is added to the existing class AdjectivePOS.</td>
</tr>
<tr>
<td>13</td>
<td>relativeAdjective</td>
<td>AdjectivePOS</td>
<td>The new individual relativeAdjective is added to the existing class AdjectivePOS.</td>
</tr>
<tr>
<td>14</td>
<td>collectiveNumeral</td>
<td>Numerals</td>
<td>The new individual collectiveNumeral is added to the existing class Numerals.</td>
</tr>
<tr>
<td>15</td>
<td>comparative</td>
<td>Degree &amp; AdjectivePOS</td>
<td>The existing individual comparative belongs to the class Degree. It should also belong to AdjectivePOS.</td>
</tr>
<tr>
<td>16</td>
<td>infinitive</td>
<td>VerbFormMood &amp; VerbPOS</td>
<td>The existing individual infinitive belongs to VerbFormMood. It should also belong to VerbPOS.</td>
</tr>
<tr>
<td>17</td>
<td>ordinalAdjective</td>
<td>AdjectivePOS &amp; NumeralsPOS</td>
<td>The existing individual ordinalAdjective belongs to AdjectivePOS. It should also belong to NumeralsPOS.</td>
</tr>
</tbody>
</table>

3 The Ontology-Based Dictionary of Russian Grammatical Forms (OntoRuGrammaForm)

In any subject area the connection of words with concepts in the form of an ontology should be based on a morpho-syntactic level. The idea turned out to be fruitful for creation of OntoRuGrammaForm. The completed experimental work made it possible
to connect words with concepts by implementing morpho-syntactic properties of the Russian language.

3.1 Description of OntoRuGrammaForm

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 (OntoRuGrammaForm). The ontology is aimed at revealing grammatical forms for the Russian words in general use.

The Ontology-Based Dictionary of Russian Grammatical Forms (OntoRuGrammaForm) contains 389,226 lemmas and 5,097,173 word forms. It is available for public use at http://ldf.kloud.one/ontorugammaform. The experience of creating the dictionary can be used for educational purposes, e.g. teaching Russian and testing knowledge of Russian.

3.2 Technical Implementation and Publication of OntoRuGrammaForm on the Web

The Open Corpora, the open corpus of the Russian language, was used as a source for OntoRuGrammaForm. 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 http://opencorpora.org/export/dict/dict.opcorpora.xsd.

The programme component of the dictionary is written in JavaScript (NodeJS), as we hold to the idea of creating and selecting the components to work with ontologies on this particular stack of technologies. We divided the technical implementation process into three blocks for convenience:

1) automatic conversion of the Open Corpora labels into the OntoLex labels;
2) for the backend we used Linked Data Fragments;
3) the client part is under development.

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/cnsttn-kndrtv/opencorpora2ontolex.

The structure of OntoRuGrammaForm conforms to the Lexicon Model for Ontologies, given in Morpho-Syntactic Description section of Community Report. As an example we use the Russian polysemantic word ‘ёж’ (‘yozh’) – ‘hedgehog’ [13]: 1) a small animal whose body is covered with sharp needle-like spines; 2) a defensive barrier of crossed girders. As we do not take meanings into account in our dictionary, these are two different words, each having its own set of morphological forms.

The description of the word, lemma, and word form relation of the word ‘ёж’ (‘yozh’) – ‘hedgehog’ in the first meaning in the Turtle format comes further.

---

4 http://opencorpora.org
5 http://linkeddatafragments.org
6 https://www.w3.org/2016/05/ontolex/#morphosyntactic-description
Fig. 1 shows the description of the word ‘ёж’ (‘yozh’) – ‘hedgehog’ in the first meaning, its lemma and three forms out of twelve.

The visualization shown in Fig.1 is implemented with the tool which is being developed now. This tool makes it possible to make federated querying to ontologies and represent query results in different forms. This kind of visualisation was specifically developed for such data types. It demonstrates convenience for representing all relations as definite groups but not as scattered vertices of a graph. This visualisation was named Terrapin (based on the name “diamond terrapin”) due to its resemblance to the Turtle format.
Fig. 1. Visualisation of relations between the morphological forms of the word ‘ёж’ (‘yozh’) – ‘hedgehog’.

4 Conclusion and Future Work
As a result of our research, we supplemented and adjusted LexInfo for the adequate description of morpho-syntactic properties of the Russian language. These supplements and adjustments are proposed as an extension to LexInfo for Russian. The supplemented and adjusted grammatical properties of Russian in LexInfo made it possible to create the Ontology-Based Dictionary of Russian Grammatical Forms (OntoRuGrammaForm) which is aimed at revealing grammatical forms of widely used Russian words. Further work will involve modeling syntactical structure of sentences with LexInfo to create a system of connecting natural language with concepts in ontologies. We also plan to create client applications for queries into OntoRuGrammaForm.

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