English-Russian WordNet for Multilingual Mappings

Sergey Yablonsky[1]

[1] St. Petersburg State University,
Volkhovsky Per. 3, St. Petersburg, 199004, Russia
{serge_yablonsky@hotmail.com}

Abstract. This paper reports about the current results of the development of the English-Russian WordNet. It describes usage of English-Russian lexical language resources and software to process English-Russian WordNet and design of a XML/RDF/OWL-markup of the English-Russian WordNet resources. Relevant aspects of the DTD/XML/RDF/OWL formats and related technologies are surveyed.

Keywords: WordNet, English-Russian WordNet, Grid, Semantic Web, RDF, OWL.

1 Introduction

The Semantic Web, a Web with the meaning**,** is often associated with specific XML-based standards for semantics, such as RDF and OWL [http://www.w3.org/RDF/, http://www.w3.org/TR/owl-features/]. If HTML and the Web made all the online documents look like one huge book, RDF, schema, and inference languages will make all the data in the world look like one huge database. One of the key promises of the Semantic Web is that it will provide the necessary infrastructure for enabling services and applications on the Web to automatically aggregate and integrate information into a sum which is greater than the individual parts. So the Semantic Web should enable users to locate, select, employ, compose, and monitor Web-based services automatically. To make use of a Web service a software agent needs a computer-interpretable description of the service, and the means by which it is accessed. An important goal for Semantic Web markup languages is to establish a framework within which these descriptions are made and shared. Web sites should be able to employ a standard ontology, consisting of a set of basic classes and properties, for declaring and describing services, while the cross-lingual ontology structuring mechanisms of OWL provide an appropriate, Web-compatible representation language framework within which to do this.

Web-compatible representation language framework today usially is based on lexical ontologies. Wordnets are cross-lingual lexical ontologies, including information on hypernyms, synonyms, polysemous terms, relations between terms, and sometimes multilingual equivalents. Wordnets are valuable resources as sources of ontological distinctions. WordNets provide a conceptual framework for

multilingual mappings in ontologies. Linking concepts across many cross-lingual lexicons belonging to the WordNet-family started by using the Interlingual Index (ILI) [2]. Unfortunately, no version of the ILI can be considered a standard and often the various lexicons exploit different version of WordNet as ILI.

At the 3rd GWA Conference in Korea there was launched the idea to start building a WordNet grid around a Common Base Concepts expressed in terms of WordNet synsets and SUMO definitions (http://www.globalwordnet.org/gwa/gwa_grid.htm). This first version of the Grid was planned to be build around the set of 4689 Common Base Concepts. Since then only three languages with essentially various number of synsets and different WordNet versions were placed in the Grid mappings (English – 4689 synsets with WN 2.0 mapping, Spanish – 15556 synsets with WN1.6 mapping and Catalan - 12942 synsets with WN1.6 mapping). But there is yet no official format for the Global WordNet Grid. So far there are just only 3 files in the specified format.

This paper reports about the current results of the English-Russian WordNet development [2, 3, 4]. It describes usage of Russian and English-Russian lexical language resources and software to process English-Russian WordNet and English-Russian WordNet Grid (4600 synsets with WN 3.0 mapping) and design of a XML/RDF/OWL-markup of the English-Russian WordNet resources. Relevant aspects of the DTD/XML/RDF/OWL formats and related technologies are surveyed.

2 Lexical Resources

2.1 Lexical Resources for English-Russian WordNet

On December 2003 our research group got license from OUP to explore and exploit for research purposes such language resources:

- Oxford Russian Dictionary;
- New Oxford Dictionary of English, 2nd Edition;
- New Oxford Thesaurus of English**.**

2.2 Language Software

For many linguistic tasks of WordNet development we use language processor Russicon that includes such main blocks:

- **System for construction and support of machine dictionaries**
  System allows receiving morphological information of the word and to build normal form of the word, shows paradigm for the word, constructs new words lexicon, constructs frequency lexicon (Fig.1).
- **Morphological analyzer and normalyzer**
  The theoretical foundation of the morphological analyzer and normalyzer program is a language-independent model of morphological analysis [6-8]. Morphological analyzer and normalyzer allows a) defining the following grammatical characteristic s of a word: part of speech, case, gender, number, tense, person, degree of comparison, voice, aspect, mood, form, type, transitiveness, reflexive, animation, b) modifying a given word to its normal grammatical form/s – lemma/s.

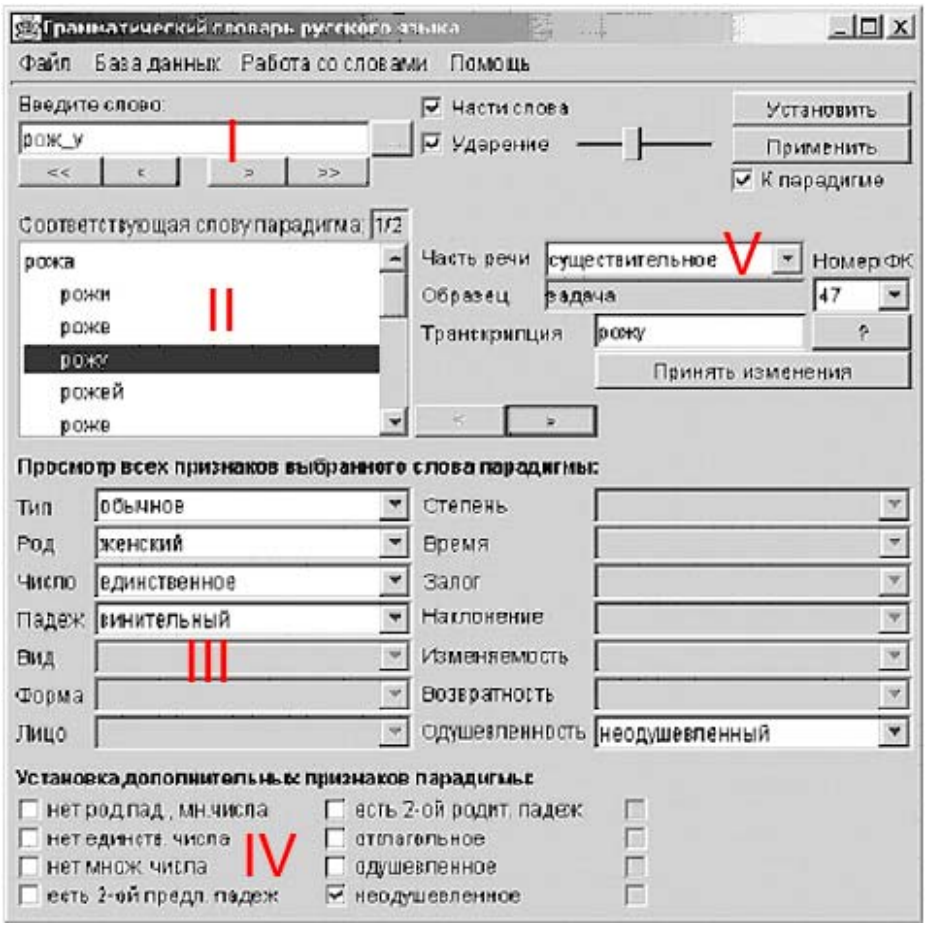

Fig.1. System for construction and support of machine dictionaries; I – word input, II – paradigm; III – input of basic grammatical features; IV – input of additional grammatical features; V – input of part of speech.

- **WordNet Editor**
  WordNet editor TenDrow was developed to help join production of Russian WordNet from above mentioned linguistic resources. It allows to join sysnsets from Thesaurus, explanatory and other dictionaries; proceed relations between synsets and words of synsets. WordNet editor is not only viewer but also a real tool for constructing and editing multiligual/monolingual WordNet. It is a database management system in which users (linguist or knowledge engineer) can create, edit and look at the English and Russian (Fig. 2).

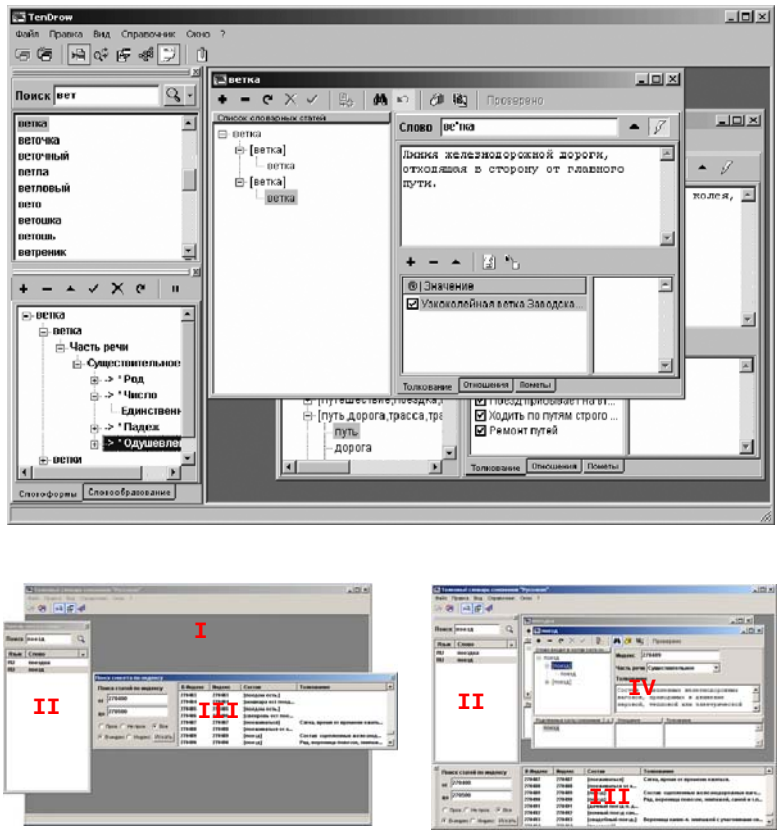

Fig. 2. TenDrow

3 English-Russian WordNet translation

Usually there were several standard variants (Fig.3, a,b,c,d) of the English-Russian WordNet and English-Russian WordNet Grid translation equivalents.

The simplest is the a variant. Approximately 24000 English-Russian synsets could be translated in such way. The hardest is d variant because such kind of translation destroys normal mapping and forms additional sub mappings. More then 15000 English synsets have no right word to word translation to Russian.

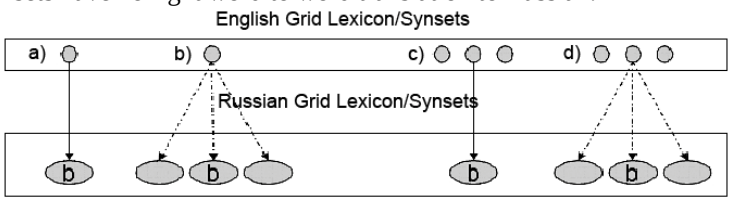

Fig. 3. Standard variants of the English-Russian WordNet [Grid] translation equivalents

Additional issues in translation could be mentioned:

- Some English Grid synsets doesn't contain the words from synsets in Example.
  Synset: **{talk of**; **talk about}** – discuss or mention
  Example: «They **spoke** of many things»
  Russian translation: obsudit'; obgovorit'; upomyanut'; kstati skazat'; kasat'sya kakoi-libo temy; kosnut'sya kakoi-libo temy.
  Synset: **{restrict}** – place under restrictions; limit access to
  Example: This substance is **controlled**
  Russian translation: ogranichivat'
  Synset: {**bring about}** – make possible
  Example: The grant **made** our research **possible**
  Russian translation: dat' vozmojnost'; obuslovit' vozmojnost'; sdelat' vozmojnym
- Some addition of the explanations in Russian translation were made in the cases when there was no any translation or when there exists only general translation not in a given sence:
  Synset: **{foot; invertebrate foot}**
  Russian translation: noga
  Russian explanation: organ peredvijeniya ili prikrepleniya u nekotoryh bespozvonochnyh
  Synset: {**soldier}**
  Russian translation: soldat
  Russian explanation: rabochaya osob' kolonii nasekomyh, prisposoblennaya k zaschite soobschestva
- Creation of new Russian synsets from English synset translation was done:
  Synset: {s**ell}** – be sold at a certain price or in a certain way
  Example: These books sell like hot cakes.
  Russian translation: {**prodavat'sya;rasprodavat'sya;sbyvat'sya}**
- Hyponomy problems: sometime no translations of English synset member exists in Russian or there were some loops in relations :
  *Synset***: {cutter, cutlery, cutting tool}** – a cutting implement; a tool for cutting.
  Hypernym: ENG20-03040079-n
  *Synset***:** {**cutting implement}** – a tool used for cutting or slicing.
  Russian translation: { kolyusche**-rejuschie orudiya}**
  Hypernym: ENG20-04279652-n
  Synset: {**edge tool}** – any cutting tool with a sharp cutting edge (as a chisel or knife or plane or gouge).
  Russian translation: { **rejuschij instrument}**
  Hypernym: ENG20-03039706-n

4 English-Russian WordNet [Grid] construction

The porting of the English-Russian WordNet was done into XML using the DTD for the XML structure from http://www.globalwordnet.org/gwa/gwa_grid.htm and the

DTD from the Arabic Wordnet: http://www.globalwordnet.org/AWN/DataSpec.html. We could use it just the same for English and Russian languages.

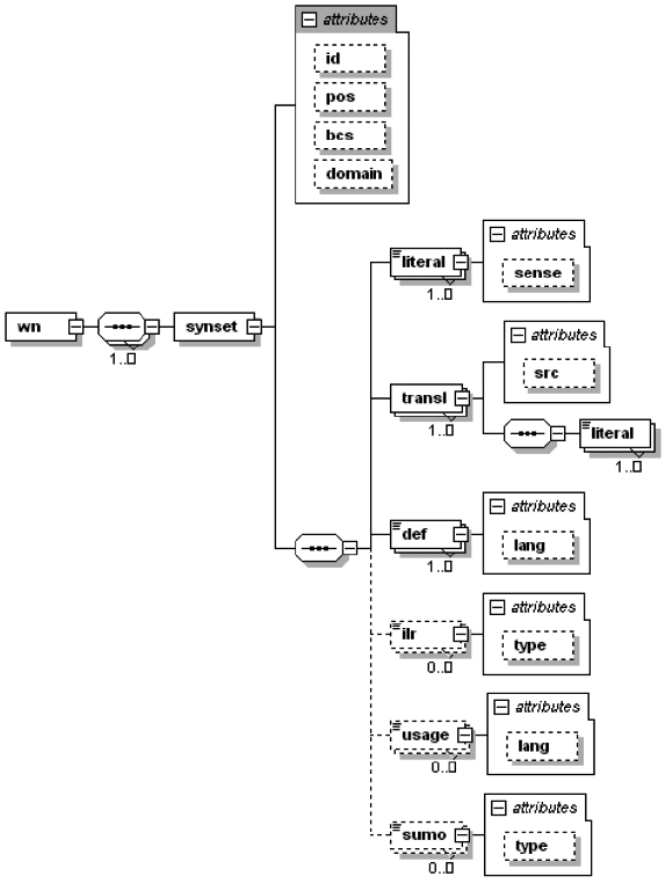

Fig.4. Standard DTD for the Russian grid XML structure

The English-Russian DTD and XML format for the English-Russian WordNet and English-Russian WordNet Grid is shown on Fig.4,5. The WordNet Task Force [9] developed a new approach in WordNet RDF conversion. The W3C WordNet project is still in the process of being completed, at the level of schema and data (http://www.w3.org/2001/sw/BestPractices/WNET/wn-conversion.html). It was used for porting of the English-Russian WordNet and English-Russian WordNet Grid into RDF and OWL.

But still there are open issues how to support different versions of WordNet in XML/RDF/OWL and how to define the relationship between them and how to integrate WordNet with sources in other languages.

5 Framework architecture for English-Russian WordNet Grid Improvement

XMLSpy 2007 and Oracle 11g were used for managing WordNet Semantic web models that provided important XML/RDF/OWL support for data modeling and editing of XML/RDF/OWL WordNet models. RDF specification defines the syntax and semantics of the SPARQL query language for RDF. SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. SPARQL also supports extensible value testing and constraining queries by source RDF graph. The results of SPARQL queries can be results sets or RDF graphs (http://www.w3.org/TR/rdf-sparql-query/).

Example. The following queries for all Synsets that contain a Word with the lexical form "bank" (http://www.w3.org/TR/wordnet-rdf/):

```
PREFIX  wn20schema: <http://www.w3.org/2006/03/wn/wn20/schema/>
SELECT  ?aSynset
WHERE   { ?aSynset wn20schema:containsWordSense ?aWordSense .
          ?aWordSense wn20schema:word ?aWord .
          ?aWord wn20schema:lexicalForm  "bank"@en-US }
```

Proposed semantic framework [8] for grid improvement is based on such main counterparts (Fig.6): RDF/OWL store; tools for information extraction; tools for Ontology Engineering Modeling Process; knowledge mining, SPAROL/SQL search and analysis tools.

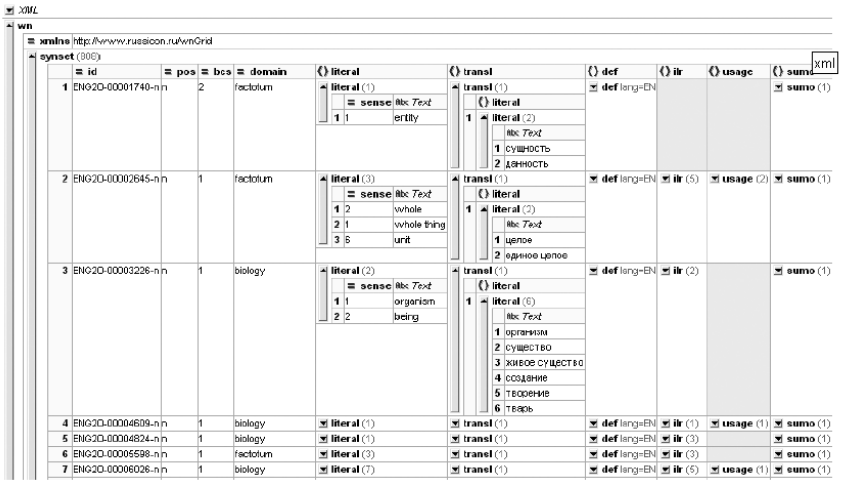

Fig.5.

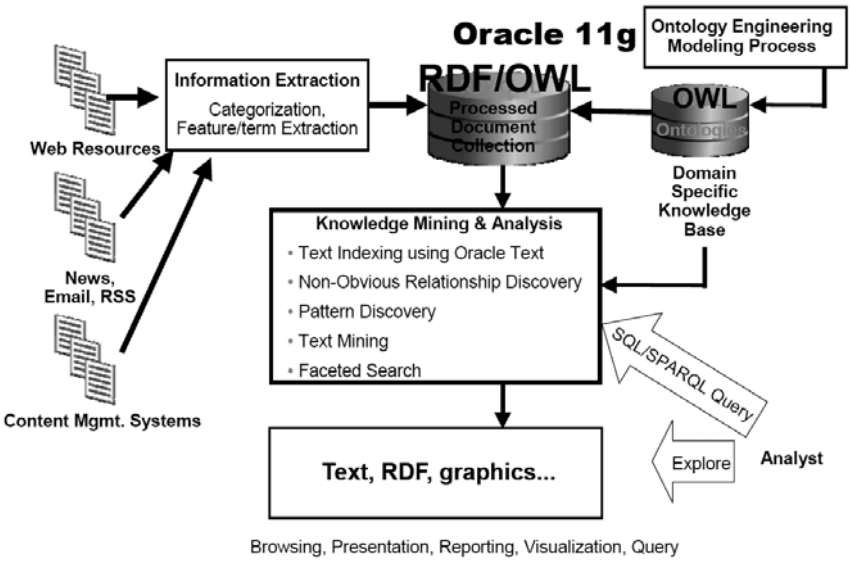

Fig.6. Semantic framework for grid improvement

6 Conclusion

Today from 117659 English WordNet synsets more then 50000 synsets have been translated from English to Russian and evaluated. Just now we are designing Web 2.0 wiki system of translation much alike as http://www.asianwordnet.org/. At the same time we have enriched English-Russian WordNet by 30000 English-Russian translations from DBpedia (http://dbpedia.org/) and LOD (http://linkeddata.org/) RDF stores.

Wordnets have been created in more then 50 of other languages http://www.globalwordnet.org/gwa/wordnet_table.htm.

This work was partly funded by The Russian Foundation for Basic Research grant 10-07-90005.

References

1. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Bradford Books (1998).
1. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dordrecht (1998).
2. Balkova, V., Suhonogov, A., Yablonsky, S. A.: Russia WordNet. From UML-notation to Internet / Intranet Database Implementation. In: Proceedings of the Second International WordNet Conference (GWC 2004), pp. 31–38. Brno (2004)
3. Balkova, V., Suhonogov, A., Yablonsky, S. A.: Some Issues in the Construction of a Russian WordNet Grid. In: Proceedings of the Forth International WordNet Conference (GWC 2008), pp. 44–55, Szeged, Hungary, January 22-25, 2008.
4. Yablonsky S. A., Suhonogov, A.: Semi-Automated English-Russian WordNet Construction: Initial Resources, Software and Methods of Translation. In: Proceedings of the Third

International WordNet Conference (GWC 2006), South Jeju Island, Korea, January 22—26 (2006).
5. Yablonsky S. A. Russicon Slavonic Language Resources and Software. In: A. Rubio, N.Gallardo, R. Castro & A. Tejada (eds.) Proceedings First International Conference on Language Resources & Evaluation, Granada, Spain (1998).
6. Yablonsky S. A. Russian Morphological Analyses. In: Proceedings of the International Conference VEXTAL, November 22-24, pp. 83-90), Venezia, Italia (1999)
7. Yablonsky S. A. Russian Morphology: Resources and Java Software Applications. In: Proceedings EACL03 Workshop Morphological Processing of Slavic Languages, Budapest, Hungary (2003).
8. Yablonsky S. A. Semantic Web Framework for Development of Very Large Ontologies, POLIBITS, Issue 39, 19–26 (2009)
9. WordNet OWL Ontology, http://www2.unine.ch/imi/page11291_en.html