# Integration of Russian Language Resources

**Serge A. Yablonsky**

Russicon Company, St.-Petersburg Transport University
Kazanskaya str., 56, ap.2
serge_yablonsky@hotbox.ru, serge_yablonsky@russicon.ru
http://www.russicon.ru

**Abstract**

In this paper we describe the creation of large scale linguistic resources for Russian language. Internet/intranet system architecture was developed to make a large volume of Russian language lexical information, corpora (texts) and knowledge base (Russian WordNet) available to the system at development and/or run time. There are four linguistic counterparts, corresponding to the major categories of lexical information developed in our system: lexicon, knowledge base, corpora and Russian language processing software.

## Introduction

Large-scale linguistic resources have been compiled and made available by organizations such as the European Linguistic Resources Association (ELRA), Linguistic Data Consortium (Comlex), Princeton University (WordNet). Systems making use of these resources can greatly accelerate the development process by avoiding the need for developers to recreate this information. In this paper we describe how we create large scale linguistic resources for Russian language. Internet/intranet system architecture was developed to make a large volume of Russian language lexical information, corpora (texts) and a large knowledge base (Russian WordNet) available to the system at development and/or run time. There are four linguistic counterparts, corresponding to the three major categories of lexical information developed in our system: lexicon, knowledge base, corpora and Russian language processing software.

## Russian lexicon

We integrate several Russian lexical resources.

The main lexicon is based on the *General Russicon Russian lexicon* which is formed from the intersection of the perfect set of such Russicon Russian dictionaries:

- *Russicon grammatical dictionary* with inflection paradigms (200.000 paradigms that produce more then 6.000.000 inflection word forms). The set of language tags consists of part of speech, case, gender, number, tense, person, degree of comparison, voice, aspect, mood, form, type, transitiveness, reflexive, animation. Lexicon consists of:
- Computer dictionary;
- Geographical names dictionary;
- Russian personal names, patronymics and surnames dictionary;
- Business dictionary;
- Juridical dictionary;
- Jargon dictionary etc.
- *The Russicon Russian explanatory dictionary*.

The dictionary gives the broad lexical representation of the Russian language of the end of the XX century. More then 100 000 contemporary entries include new words, idioms and their meanings from the language of the Eighties-Nineties. The dictionary is distinguished by its complete set of entry word characteristics, clear understandable definitions, its guidance on usage. All dictionary information for entries is structured in more then 60 attributes:

  - entry word;
  - multiple word entries;
  - usage notes;
  - precise, contemporary definitions;
  - derivations;
  - example sentences/citations;
  - idioms etc.

- *The Russicon Russian thesaurus* (set of 14.000 Russian synsets). Synonym list plus word list containing approximately 30 000 normalized entry words with inflection paradigms.
- *The Russicon Russian Orthographic dictionary***.**

Additionally we use some entry words from most popular Russian dictionaries from 19th century up to the end of 20th (print and available through Internet). For example:

- *Russian Explanatory Dictionary* (Efremova T.F., 2001 – 136.000 entry words) for improvement of the Russian WordNet structure;
- *The Russian Semantic Dictionary* (ed. Shvedova N.Y.,1998,2000, vol.1, 2 – 39.000 + 40.000 entry words);
- *The Explanatory Ideographical Dictionary of Russian Verbs* (Babenko L.G., 1999 – 25000 entry words) for improvement of the Russian hyponomy/hyperonymy and meronomy/holonymy relations.
- *The Russian Language Antonyms Dictionary* (L'vov M.R., 2002– 3200 entry words) for improvement of the Russian antonomy relations

## Knowledge base

The knowledge base is found on Russian WordNet (http://www.pgups.ru/WebWN/wordnet.uix), a machine-readable hierarchical network of concepts which is in a stage of development. Concepts in WordNet do not have names - they are just sets of words (called synsets). The Russian WordNet concepts correspond to real-world entities and phenomena in terms of which people understand the meanings of words. Our Russian knowledge base is currently concerned with the concepts corresponding to nouns, adjectives, verbs and adverbs.

The current status of the Russian WordNet is represented with such statistics of synsets in the first version, February 2004 update (see: Table 1, Table 2).

| Total | Noun | Verb | Adj | Adv |
|---|---|---|---|---|
| 111904 | 45424 | 29421 | 21317 | 5347 |

Table 1: Russian WordNet word report.

| WordCnt | Total | Noun | Verb | Adj | Adv |
|---|---|---|---|---|---|
| 1 | 120102 | 53791 | 29471 | 25811 | 5181 |
| 2 | 9289 | 2845 | 4893 | 1286 | 196 |
| 3 | 2860 | 974 | 1342 | 391 | 135 |
| 4 | 1847 | 555 | 908 | 268 | 99 |
| 5 | 1233 | 359 | 584 | 200 | 80 |
| 6 | 942 | 262 | 443 | 162 | 66 |
| 7 | 655 | 186 | 328 | 96 | 41 |
| … | … | … | … | … | … |
| Total | **138947** | **59395** | **39036** | **28563** | **5955** |

Table 2: Russian WordNet synset report.

The list of semantic relations in WordNet.ru is based mostly on Princeton WordNet Lexical and Conceptual Relations, and EuroWordNet Language-Internal Relations.

Main relations between synsets:

- hyponymy/hyperonymy,
- antonymy,
- meronomy/holonymy etc.

Main relations between members of synsets:

- synonymy,
- antonymy,
- derivation synonymy,
- derivation hyponymy.

Two last relations are relations between aspect pairs and between neutral words and their expressive derivatives etc.

We produce inflection paradigm for every input word. This gives us possibility to output Russian WordNet synsets not only for lemma of input word, but for any inflection form of input word. It is important because Russian is highly inflection language.

Two complementary approaches were devised in EuroWordNet to build local wordnets from scratch:

- The merge approach: building taxonomies from monolingual lexical resources and then, making a mapping process using bilingual dictionaries
- The expand approach: mapping directly local words to English synsets using bilingual dictionaries.

The merge and expand approaches are both present in our Russian WordNet construction process from the beginning.

We are really building taxonomies using Russian lexical resources mentioned above.

At the same time we use the expand approach for direct mapping of many words from English WordNet to Russian and vise verse. Only this approach is used for some English proper and geographical names.

We present the open UML-specification and new pilot database management system on Oracle 9i DBMS for efficient storage and retrieval of various kinds needed to process English-Russian WordNet. Relevant aspects of the UML/ER data models and related technologies are surveyed. Bilingual WordNet system could be easily expanded in a real multiligual system (Balkova V., Suhonogov A., Yablonsky S., 2004).

## Russian corpora

Corpus composition is based on wide representation of Russian literature, critics, philosophy, religion, newspapers, memoirs, law, business, computers, historical documents, translations, folklore, Internet literature, "underground" literature, some transcriptions of everyday casual conversation, radio broadcasts, meetings, interviews and discussions etc published after 1975 and selected so as to reflect the present day written Russian. Texts are received mostly in electronic form from Internet and CD–resources, but some are taken from printed editions (Yablonsky S., 1998). Resources architectural specification provides software developers, and content developers with a common reference for interoperable text manipulation on the World Wide Web, building on the Universal Character Set, defined jointly by the Unicode Standard and ISO/IEC 10646. For normalization and string identity matching, see the companion document Character Model for the World Wide Web 1.0 (www.w3.org).

We present the open UML-specification of corpora system that could be expanded in future by community of language and speech software and resources developers (Yablonsky S., 2003-1). Reusability of linguistic and corpora manipulation business services is achieved by usage of a widely accepted set of notation standards for corpus-based work in natural language processing applications. System could be easily adapted to different software platforms (Java and Oracle) and the necessities of other languages (Unicode), making the whole system mobile. The only language-specific resources are a large-scale morphosyntactic dictionary plus POS tagger.

Today the Web is clearly a multilingual corpus (Kilgarriff A., Grefenenstette G., 2003). The Web contains enormous quantities of text in numerous languages and in Russian particular. So we started pilot project “Monitoring of Russian Web” (Yablonsky S., 2000; Yablonsky S., 2003-1).

The main goal is to develop a “linguistic” search engine that could provide

- enough instances (no number limitations, only time limitations),
- enough context (full sentence minimum);
- search to be specified according to linguistic criteria such as citation form for a word, word class, word lemma, searches are specified in terms of lemmas, noun phrases and grammatical relations rather then strings,
- reliable statistics.

The future aim is dynamic parsing of the Web.

## Language software

For many linguistic tasks of integration development we use such parts of language processor Russicon (Yablonsky S.A. 1998, 1999, 2003): system for construction and support of machine dictionaries, morphological analyzer and normalyzer, word sense disambiguation.

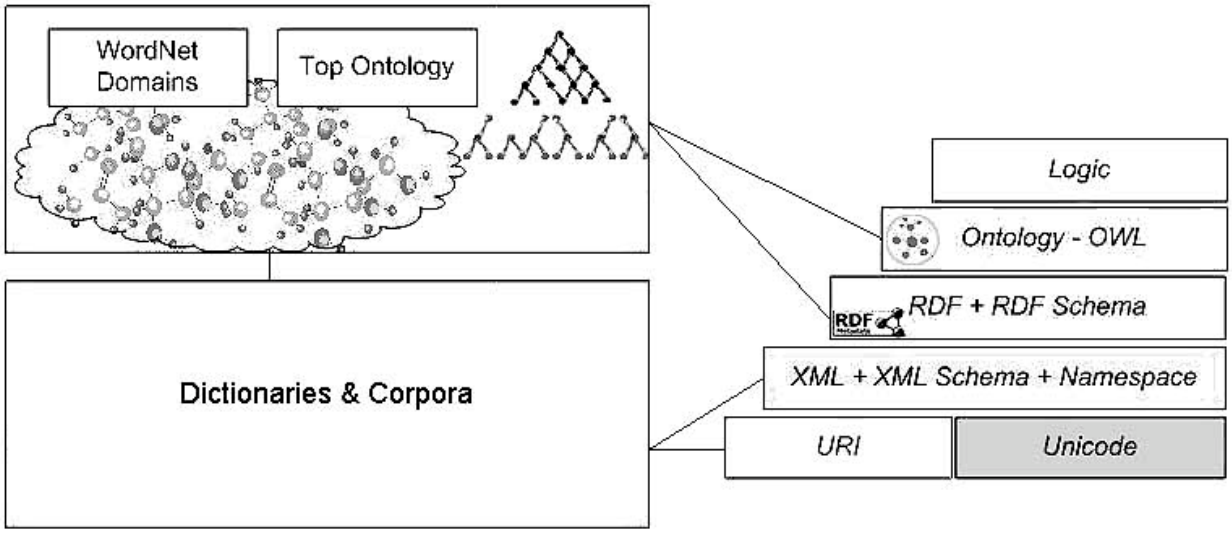


Figure 1: Usage of Semantic Web standarts for Russian language resourses

## Integration of resources

Information from Russian resources is linked together via the lemma. Russian lemmas can have more than one meaning. In this case the relevant meaning is determined and isolated from the other meanings by word sense disambiguation.

Integration of mentioned above counterparts is made by such steps:

- Development of all-lexicon of all words extracted from all resources.
- Development of Oracle/Java database system (Yablonsky S.A., 2003-1) to interconnect all man parts by all-lexicon lemmas (paradigms) with the help of language processor (morphological analyzer, normalyzer, syntactic/semantic analyzers - Yablonsky S.A., 1998, 2003-2,).
- Annotating main corpora by the help of word sense disambiguation.

The Web can reach its full potential only if it becomes a place where data can be shared and processed by automated tools as well as by people. The Semantic Web is a vision: the idea of having data on the web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications. In 2004 the World Wide Web Consortium released the Resource Description Framework (RDF) and the OWL Web Ontology Language (OWL) as W3C Recommendations for Semantic Web. RDF is used to represent information and to exchange knowledge in the Web. OWL is used to publish and share sets of terms called ontologies, supporting advanced Web search, software agents and knowledge management. Our usage of Semantic Web standards is shown on Figure 1.

## Conclusion

We present basic resources and software for content-based language-technologies within and across the Russian language. It enables different forms of text indexing and retrieval and a direct benefit from the integrated linguistic resources in:

- information-acquisition tools,
- authoring tools,
- language-learning tools,
- translation-tools,
- summarizers,
- semantic web.

## References

Kilgarriff A., Grefenenstette G. (2003). Introduction to the Special Issue on the Web as Corpus. Computation Linguistics, Special Issue on the Web as Corpus.

Yablonsky S. (1998). Russicon Slavonic Language Resources and Software. In: A. Rubio, N. Gallardo, R. Castro & A. Tejada (eds.) Proceedings First International Conference on Language Resources & Evaluation, ( pp. 1141–1147), Granada, Spain.

Yablonsky S. (1999). Russian Morphological Analyses. In: Proceedings of the International Conference VEXTAL, November 22–24 1999, (pp. 83 – 90), Venice, Italy.

Yablonsky S. (2000) Russian Monitor Corpora: Composition, Linguistic Encoding and Internet Publication. In Proceedings Second International Conference on Language Resources & Evaluation, Athens, Greece.

Yablonsky S. (2002) Corpora as Object-Oriented System. From UML-notation to Implementation. In Proceedings Third International Conference on Language Resources & Evaluation,Las Palmas,Canary Islands-Spain.

Yablonsky S. (2003-1). The Corpora Management System Based on Java and Oracle Technologies. In Proceedings 10th conference of the European Chapter of the Association for the Computational Linguistics, Budapest, Hungary.

Yablonsky S. (2003-2). Russian Morphology: Resources and Java Software Applications. In Proceedings EACL03 Workshop Morphological Processing of Slavic Languages, Budapest, Hungary.

Balkova V., Suhonogov A., Yablonsky S. (2004). Russia WordNet. From UML-notation to Interne/Intranet Database Implementation. In Proceedings of the Second International WordNet Conference, GWC 2004, Brno, Czech Republic, 2004, pp. 31-38.