Russian WordNet

From UML-notation to Internet/Intranet Database Implementation

Valentina Balkova[2], Andrey Sukhonogov[1], and Sergey Yablonsky[1,2]

[1] Petersburg Transport University, Moscow av., 9, St.-Petersburg, 190031, Russia,
Email: ASukhonogov@rambler.ru
[2] Russicon Company, Kazanskaya str., 56, ap.2, 190000, Russia
Email: v_balk@front.ru,serge_yablonsky@hotmail.com,root@russicon.ru

Abstract. This paper deals with development of the first public web version of Russian WordNet and future parallel English-Russian and multiligual web versions of WordNet. It describes usage of Russian and English-Russian lexical language resources and software to process WordNet for Russian language and design of a database management systems for efficient storage and retrieval of various kinds of lexical information needed to process WordNet. Relevant aspects of the UML data models, XML format and related technologies are surveyed. The pilot Internet/Intranet version of described system based on Oracle 9i DBMS and Java technology is published at: http://www.pgups.ru/WebWN/wordnet.uix.

1 Introduction

This paper attempts to introduce results of an ongoing project of developing of the first public web version of Russian WordNet and future parallel English-Russian and multiligual web versions of WordNet. English-Russian parallel WordNet resources and software implementation for building parallel multilingual lexical database based on Princeton WordNet are introduced. The goal of database management system development is to build a multilingual (monolingual Russian now and bilingual English-Russian and multilingual in future) lexical database of wordnets for Russian language (WordNet.ru), which are structured along the same lines as the Princeton WordNet for English language. WordNet.ru contains information about nouns, verbs, adjectives and adverbs in Russian and is organized around the notion of a synset. The WordNet.ru represents basic resources for content-based language-technologies within and across the Russian and English languages. It will enable a form of multilingual text indexing and retrieval, a direct benefit from the multilingual semantic resource in:

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

The objectives of this project are not unique. Several analogous projects have been carried out to different stages (EuroWordNet, BalkanNet etc.) but there is no public web realization of Russian WordNet yet.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 31–38.

We have been implementing a combination of manual and automatic techniques. Today there are several WordNet viewers: the Princeton viewer and EuroWordNet viewer/editor (VisDic). The limitations of these popular WordNet tools for Russian WordNet design stimulate our development of Russian WordNet editor and Multilingual WordNet editor based on Oracle database management system.

The paper discusses the complete process of building and managing of monolingual Russian and parallel English-Russian version of WordNet database management system, including the development of UML/ER-specifications, architecture and examples of actual implementations of DBMS tools. The system is implemented using DBMS Oracle9i Release 2 and Java technology.

2 Lexical Resources for Russian WordNet

We use several Russian lexical resources. Russicon company has such main counterparts (Yablonsky S. A., 1998, 2003) for English-Russian and Russian WordNet development:

- *The General Russicon Russian lexicon* which is formed from the intersection of the perfect set of Russicon Russian grammatical dictionaries with inflection paradigms (200,000 paradigms that produce more then 6,000,000 inflection word forms). Lexicon consists of:
    - Russian basic grammatical dictionary;
    - 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.*

All dictionaries are implemented as text-files and as compressed linguistic databases connected to the Russicon language processor. *Text-files* of grammatical dictionaries contain normalized entry words (lemmas) with hyphenation and inflexion paradigm plus grammatical tags for each word of paradigm. 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. For thesaurus and explanatory dictionary we have two or more text – files, one always containing inflexion paradigms of all words of the dictionary. Formats of files are plain text and HTML.

We also use several print Russian dictionaries:

- a version of the new monoligual *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) and *The Explanatory Ideographical Dictionary of Russian Verbs* (Babenko L. G., 1999 – 25000 entry words) for improvement of the Russian WordNet hyponomy/hyperonymy and meronomy/holonymy relations;
- *The Russian Language Antonyms Dictionary* (L'vov M. R., 2002 – 3200 entry words) for improvement of the Russian WordNet antonomy relations.

3 English-Russian WordNet

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 approach is present in our Russian WordNet construction process from the beginning. We are really building taxonomies using Russian lexical resources mentioned above. After our first version will be finished we plan making mapping using bilingual dictionaries.

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

4 The Current Status of the Russian WordNet

The statistics of synsets in the first version of WordNet.ru are displayed in Table 1.

We plan to include additionally 10,000 Russian local proper and geographic names in the first version.

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. Main relations between members of synsets: synonymy, antonymy, derivation synonymy,

Table 1. Statistics of synsets in the first version of WordNet.ru

Russian WordNet Word Report					
Total	Noun	Verb	Adj	Adv	Other
111749	44751	27997	20736	4997	13268

Synset report						
WordCnt	Total	Noun	Verb	Adj	Adv	Other
1	120549	53137	29351	25299	4976	7786
2	12825	3355	7077	1635	188	570
3	3637	1011	1675	378	121	452
4	2193	574	920	253	89	357
5	1424	351	581	186	78	228
6	1121	258	428	148	67	220
7	791	184	311	89	45	162
8	565	128	239	58	37	103
9	443	72	186	62	26	97
10	305	55	124	45	16	65
...	...	...	...	...	...	...
68	2	0	0	0	1	1
Total	144980	59294	41403	28316	5718	10249

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. The number of all inflections is approximately 5,000,000. 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.

5 Language Software

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

6 WordNet Conceptual Model

6.1 UML Model Design

Today Unified Modeling Language (UML) defines a standard notation for object-oriented systems (Booch G., Rumbaugh J., and Jacobson I., 1998). Using UML enhances communication between linguistic experts, workflow specialists, software designers and other professionals with different backgrounds. At the same time UML diagrams are widely used for realation data base design (for example in Rational Rose).

The core part of Russian WordNet UML model includes **SYNSET, WORD, IDIOM, EXPLANATION** entities (Figure 1). For **SYNSET** entity such attributes are defined:

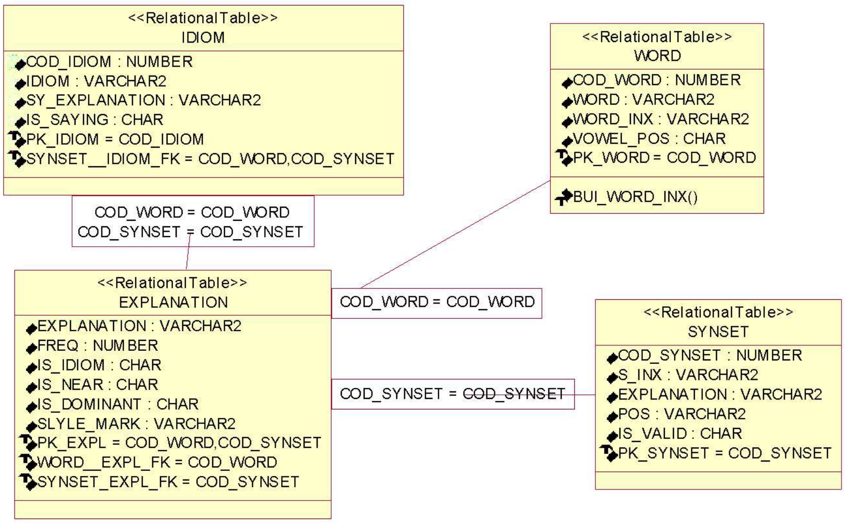

Fig. 1. Core part of Russian WordNet UML model

- COD_SYNSET (internal database synset identifier);
- S_INX (index) – unique synset identifier; it could be defined by user while working with thesaurus;
- EXPLANATION – synset explanation;
- POS – grammatical information;
- IS_VALID – validation flag.

For **WORD** entity such attributes are defined:

- COD_WORD (word identifire) – internally used database primary key;
- WORD (word code);
- WORD_INX (word index) – internally used database key for word search;
- VOWEL_POS (stress vowel) – stress position information string; up to 4 stresses in one word could be fixed;

Synset includes one or more words (lemmas) and one word could be included in more then one synset.

Entity **EXPLANATION** is used for storing information about meaning of the word. For it such attributes are defined:

- EXPLANATION (word meaning) – natural language word meaning description;
- IS_IDIOM – idiom identification flag (is true for idiom);
- IS_NEAR – near word identification flag;
- IS_DOMINANT – synset dominant word identification flag;
- STYLE_MARK.

For entity **IDIOM** such attributes are defined:

- COD_ IDIOM – internally used database primary key;
- IDIOM;
- SY_EXPLANATION – natural language idiom meaning description;
- IS_SAYING – saying identification flag.

In Russian WordNet model all types of WordNet relations between synsets are realized. Even more, there are no limitations on the type of relations between synsets. The semantics and number of relations is user defined. For that purpose user is given the so-called *sematic/type relation constructor*. Types of relation are devided into two main groups: *hierarchic (symmetric) and not hierarchic (symmetric and not symmetric).* In Russian WordNet model we plan to develop domain WordNets.

6.2 ER Model Design

At the same time ER (Entity-Relation) models are also very popular in relational data base design. Figure 2 presents the whole ER model of Russian WordNet.

7 Main Steps of Russian WordNet Development

The development process of Russian and English-Russian WordNet development could be devided into two main steps.

The first step ends by production of the first version of Russian WordNet with the number of word inputs more then 100,000. The exact numbers could be found in Section 4. For construction of Russian WordNet we developed Russian WordNet editor.

The second step ends by development of English-Russian version of WordNet. For that purpose we developed Multilingual WordNet Editor.

7.1 Russian WordNet Editor

Russian WordNet editor was developed to help 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.

It is a database management system in which users (linguist or knowledge engineer) can create, edit and view Russian WordNet. From a monolingual point of view they can work with any monolingual WordNet (for us – Russian) with its internal semantic relationships.

7.2 Multilingual WordNet Editor

We designed multiligual WordNet editor (beta version) that includes definition of the relations, the common data structure, the shared ontology, the Inter-Lingual-Index and the comparison option, Russian so-called Base Concepts (the Base Concepts are the major building blocks on which the other word meanings in the wordnets depend).

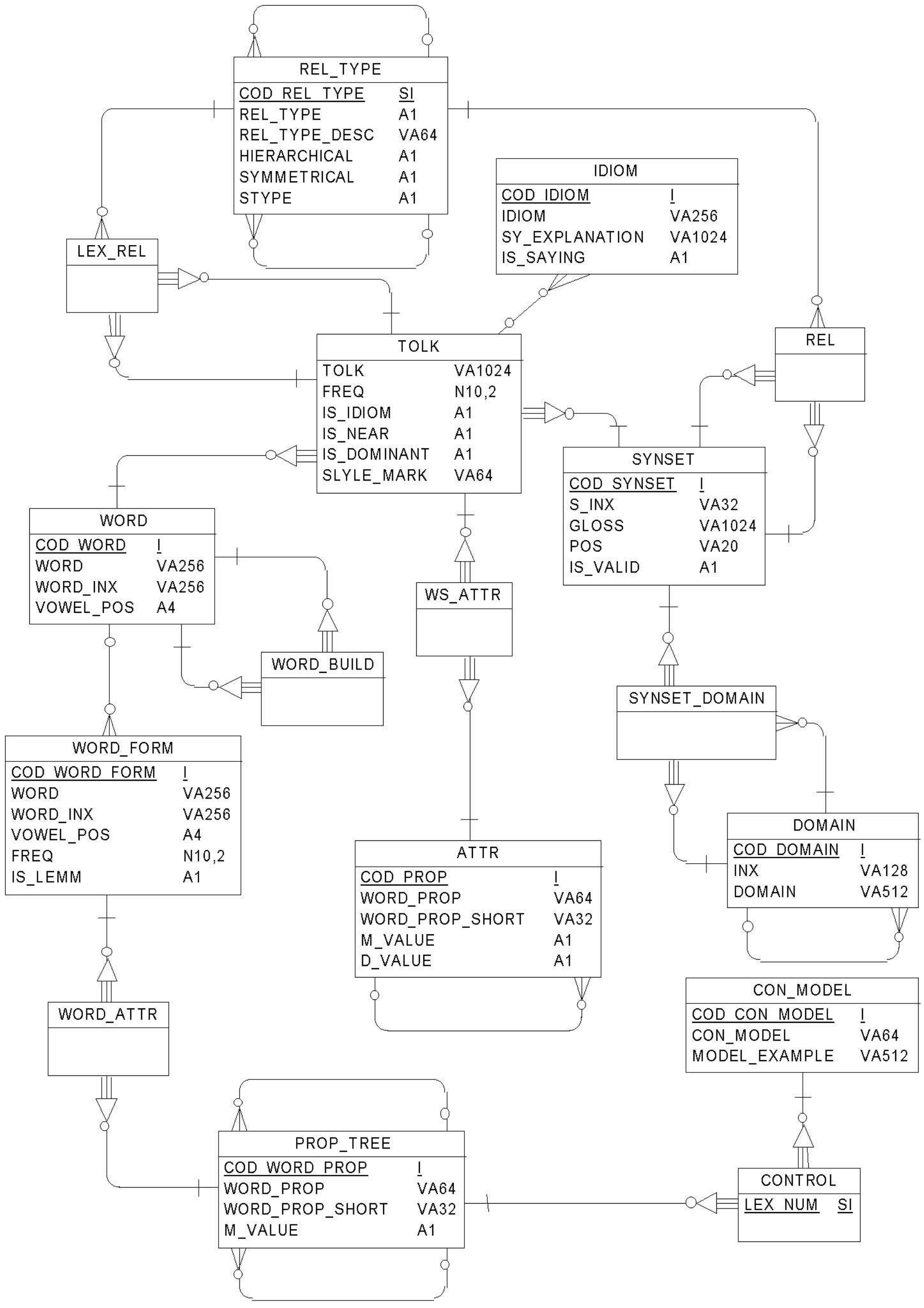

Fig. 2. Russian WordNet ER model

8 Internet Viewer

The pilot Internet version of described system based on Oracle 9i DBMS and Java technology is published at: `http://www.pgups.ru/WebWN/wordnet.uix`.

Our Internet/Intranet WordNet viewer is a database management system in which users (linguist or knowledge engineer) can look at the Russian and English WordNet databases.

9 Conclusion

We present the open UML-specification and new pilot database management system on Oracle 9i DBMS for efficient storage and retrieval of various kinds of lexical information 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.

References

1. Babenko L. G. – ed., 1999. Explanatory Ideographical Dictionary of Russian Verbs. – Moscow, Ast-Press.
2. Booch, G., Rumbaugh, J., and Jacobson, I., 1998. The Unified Modeling Language user guide, Addison-Wesley.
3. Efremova T. F., 2001. Novij Slovar Russkogo Yazika [Новый словарь русского языка], v.1,2. – Russian Language.
4. Fellbaum C. WordNet: an Electronic Lexical Database. MIT Press, Cambridge, MA,1998.
5. Lyons J. Semantics. (2 vol.) London and New York, 1977.
6. L'vov M. R., 2002. The Russian Language Antonyms Dictionary. – Moscow, Ast-Press.
7. Miller G. et al. Five Papers on WordNet. CSL-Report, vol.43. Princeton University, 1990.
8. `ftp://ftp.cogsci.priceton.edu/pub/wordnet/5papers.ps`.
9. Prószéky, Gábor & Márton Miháltz, 2002. Semi-automatic Development of the Hungarian WordNet. LREC-2002, Las Palmas, Spain.
10. Shvedova N. Y. – ed., 1998. Russian Semantic Dictionary, vol.1. – Moscow, Azbukovnik.
11. Shvedova N. Y. – ed., 2000. Russian Semantic Dictionary, vol.2. – Moscow, Azbukovnik.
12. Vossen, P. EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht.
13. Kluwer, 1998.
14. Yablonsky S. A., 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, Granada, Spain.
15. Yablonsky S. A. (1999). Russian Morphological Analyses. In: Proceedings of the International ConferenceVEXTAL, November 22–24 1999, (pp. 83–90), Venezia, Italia.
16. Yablonsky S. A. (2003). Russian Morphology: Resources and Java Software Applications. In: Proceedings EACL03 Workshop Morphological Processing of Slavic Languages, Budapest, Hungary.