

Construction of a Russian WordNet Grid

Valentina Balkova³, Sergey Yablonsky^{1,2,3}

¹Graduate School of Management, Information Technologies in Management Department,
St. Petersburg State University,

²Petersburg Transport University, Information Systems Department,

³Russicon Company

v_balk@front.ru, serge_yablonsky@hotmail.com

Computational lexicons (CL) provide machine understandable word knowledge. Semi-automatic integration and enrichment of large-scale multilingual lexicons like WordNet [1] is used in many computer applications. The three core concepts in WordNet are the synset, the word sense and the word. Words are the basic lexical units, while a sense is a specific sense in which a specific word is used. Synsets group word senses with a synonymous meaning, such as {car, auto, automobile, machine, motorcar} or {car, railcar, railway car, railroad car}. There are four disjoint types of synset, containing exclusively nouns, verbs, adjectives or adverbs. There is one specific type of adjective, namely an adjective satellite.

Furthermore, WordNet defines seventeen relations, of which

- ten between synsets (hyponymy, entailment, similarity, member meronymy, substance meronymy, part meronymy, classification, cause, verb grouping, attribute);
- five between word senses (derivational relatedness, antonymy, see also, participle, pertains to);
- “gloss” (between a synset and a sentence);
- “frame” (between a synset and a verb construction pattern).

Linking concepts across many 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. As alternative another possible solution can use the DTD from the Arabic WordNet: <http://www.globalwordnet.org/AWN/DataSpec.html>.

This paper reports about the result of the development of the Russian WordNet Grid [3]. It describes usage of Russian and English-Russian lexical language resources and software to process WordNet Grid for Russian language (4600 synsets with WN 2.0 mapping) and design of a XML/RDF/OWL-markup of the grid resources. Relevant aspects of the DTD/XML/RDF/OWL formats and related technologies are surveyed.

Several Russian lexical resources were used for the Russian WordNet Grid [4]. We've done porting of the original English and Russian WordNet [3] 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>. The standard DTD for the Russian grid XML structure and the English/Russian XML format for the Grid was developed. The grid of English and Russian local WordNets is realized as a virtual repository of XML databases accessible through web services. Basic services devoted to the management of the actual versions of Princeton and Russian WordNets. Unfortunately, no version of the grid can be

considered a standard because the various grids exploit different versions of WordNet, have different numbers of entries and there is no mappings of the multilingual grids on new versions of WordNet. We've done porting of the original English and Russian WordNet Grid into RDF and OWL [4].

References

1. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Bradford Books (1998).
2. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dordrecht (1998).
3. 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)
4. Balkova, V., Suhonogov, A., Yablonsky, S. A.: Some Issues in the Construction of a Russian WordNet Grid. In: Proceedings of the Fourth International WordNet Conference (GWC 2008), pp. 44–55, Szeged, Hungary, January 22-25, 2008.