Serbian Wordnet for biomedical sciences

Sanja Antonic
University library ”Svetozar Markovic” University of Belgrade, Serbia
antonic@unilib.bg.ac.yu

Cvetana Krstev
Faculty of Philology, University of Belgrade, Serbia
cvetana@matf.bg.ac.yu

INFORUM 2008: 14th Conference on Professional Information Resources
Prague, May 28-30, 2008

Abstract: Wordnet is a lexical semantic network having as nodes the sets of synonymous word (synsets) which are linked by various semantic relations. The first Wordnet built was the so-called Princeton Wordnet consisting today of approximately 140,000 synsets. Due to its remarkable size and successful inclusion in various computer-based applications it is considered as a de facto standard. In this paper we present the new electronic information resource Serbian Wordnet (SWN) and its initial development. Serbian Wordnet is based on the same principles as Princeton Wordnet (PWN), which means that it has not only inherited its basic structure but also the additional information associated to PWN. The development of Serbian Wordnet started with Balkanet project which was funded by European Commission from (2001-2004). The main goal of the project was development of wordnets for the Balkan languages: Bulgarian, Greek, Romanian, Serbian, Turkish and Czech and their alignment with PWN. The development of Serbian Wordnet continues after Balkanet with its enhancement in various domains. We describe how various traditional and Internet resources are being used for the enhancement of SWN in domain of biomedicine and how we see the possible practical usage of SWN in information retrieval this domain.
Keywords: lexical semantic network, Serbian WordNet, biomedicine.

Wordnet (WN) is a semantic network of concepts represented by synsets – sets of synonymous words or more precisely different Part of Speech-PoS (nouns, verbs, adjectives, adverbs). It is based on a grouping of synonyms into synsets which are representing network nodes. Each node are interconnected by one or more arcs which describe particular semantic relations (hyperonymy, hyponymy, antonymy etc.) [1] Generally, every synset is accompanied by a definition (gloss) and examples of usage that specify the meaning of the concept represented by the synset.

Initially, Wordnet was developed by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller for the English language. WordNet is online lexical reference system and its structure makes it a useful tool for computational linguistics and natural language processing. The Princeton WordNet is the standard model for the development of

more than 50 monolingual wordnets for more than 40 languages all over the world and one of them is Serbian Wordnet (SWN). Serbian Wordnet is based on the same principles as Princeton Wordnet (PWN), which means that it has not only inherited its basic structure but also the additional information associated to PWN, most notably information about the domain and SUMO category assigned to every synset.

Possible applications of Wordnets are:

- machine translation;
- Wordnets can be used as a new type of dictionary according to synsets (synonymy relation);
- information extraction, allows to follow semantic relations in text, and exploit multilinguality,
- useful with web browsers,
- word sense discrimination – as a data resource for sense recognition, knowledge representation,
- inference relying on word meanings, relations to Semantic web etc.

EuroWordNet project added a completely new component to the Princeton WordNet, namely multilinguality. The EuroWordNet (EWN) (LE-2 4003 & LE-4 8328), which started in March 1996 and ended in June 1999, extended the PWN approach with the multilingual dimension adding an Inter-Lingual Index to which all the monolingual wordnets for the languages represented in the project were aligned. The languages represented in EWN were Dutch, Italian, Spanish, German, French, Czech, Estonian and obviously English.

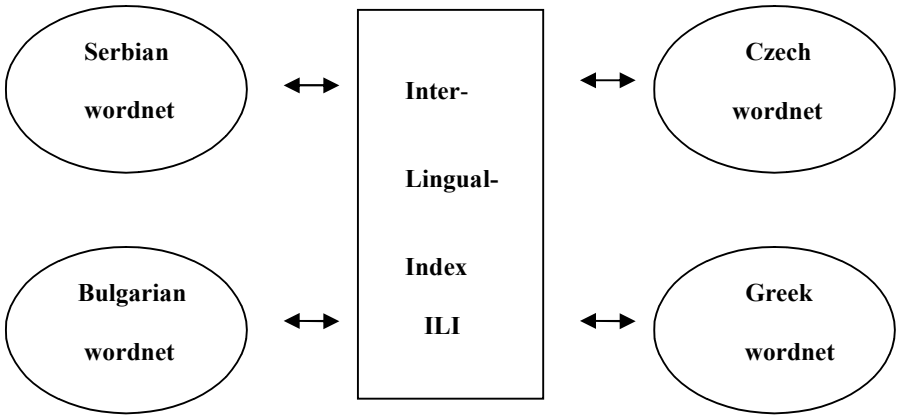


Figure 1. Inter-Lingual Index

A few years after EuroWordNet, the BalkaNet project continued and intensified development of Wordnet. BalkaNet was an EC funded project (IST-2000-29388) that started in September 2001 and finished in August 2004. It aimed at developing aligned wordnets for the following Balkan languages:

Bulgarian, Greek, Romanian, Serbian, Turkish and to extend the Czech wordnet previously developed in the EuroWordNet project. BalkaNet project had in so far delivered many useful results in the fields of both Computational Lexicography and Natural Language Processing (NLP). In addition to the many tools and sub-resources developed by the separate partners, a major achievement is the maturation of the VisDic tool. The VisDic database is widely used not only in the BalkaNet project but also outside the project. It is currently the standard database for developing wordnets, enabling formal and conceptual consistency checking and supporting various formats, but in Balkanet project a common format was XML
[2].

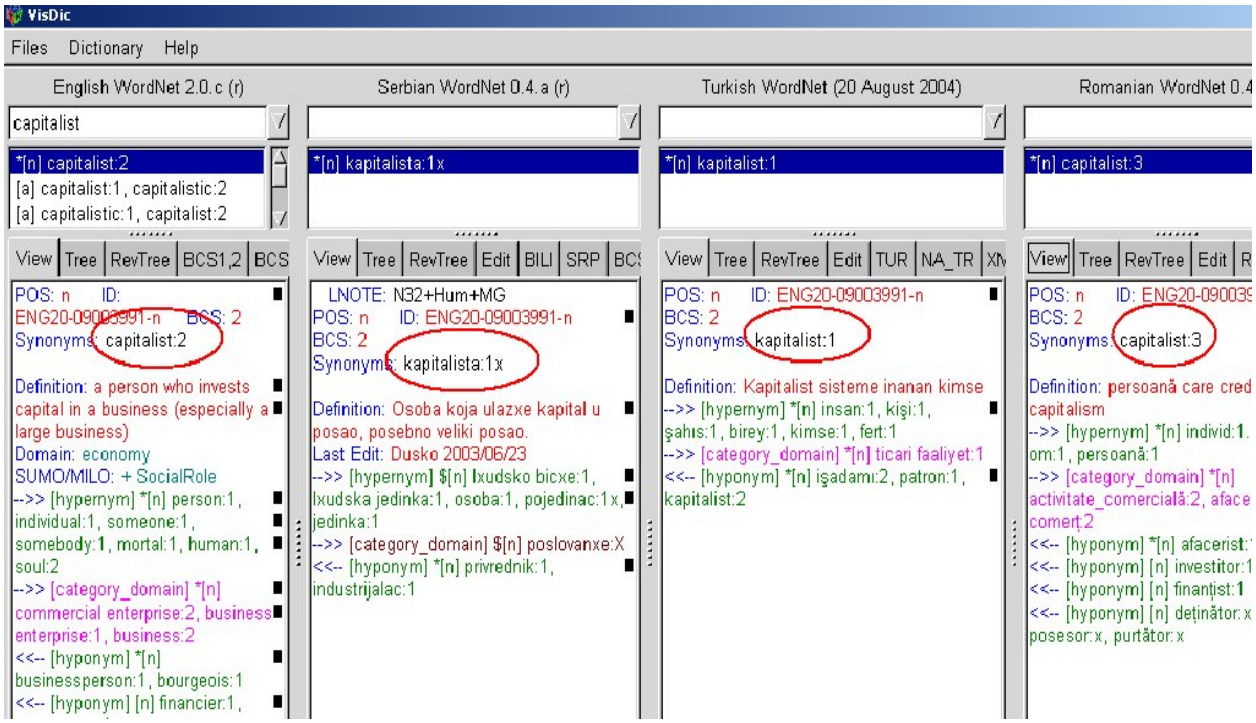


Figure 2. VisDic - multilingual viewer and editor

Main goals of Balkanet were: at least 8000 synsets per partner, maximal inter-lingual overlap (> 80%), building tools to efficiently exploit the multilingual wordnet (sum of the ILI – based aligned monolingual wordnets), development of free software for use and management of the BalkaNet wordnets, building various application (WSD, intelligent document indexing, CLIR, etc.)

Design principles were: ensuring as much as possible compatibility with the EWN approaches (e.g. unstructured ILI based on PWN), synset structuring (relation) inside each wordnet (a lot of redundancy, but much more powerful), keeping up with PWN developments, defining a reusable methodology for data acquisition and validation, linguistically motivated (reference language

resources, with human experts actively involved in all decision making and validation), minimizing the development time and costs.

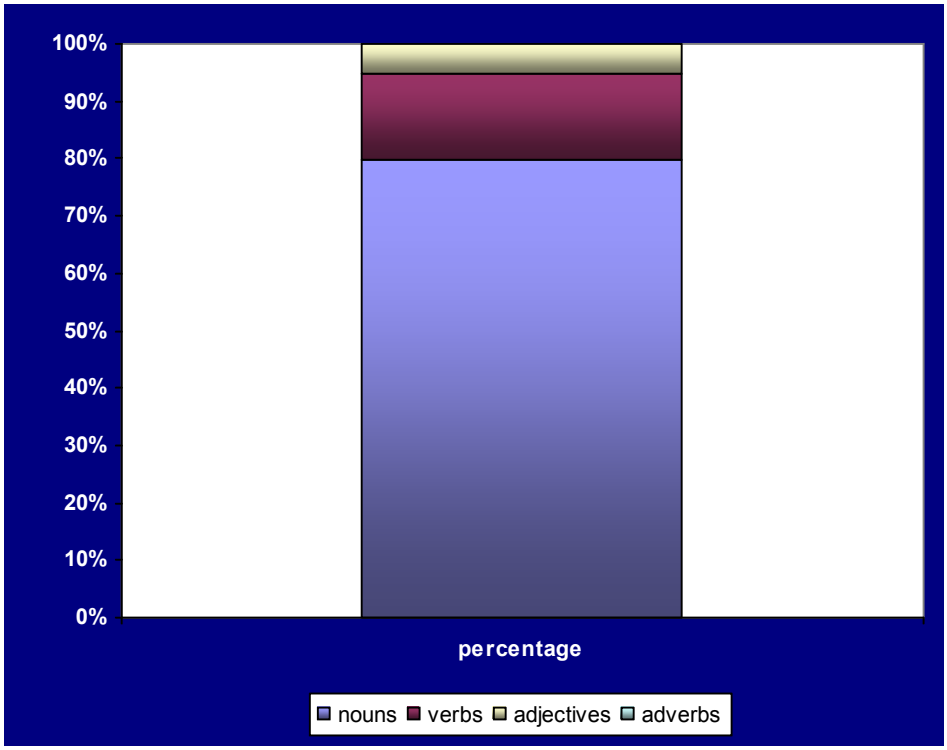
The next task was maximization of the cross-lingual coverage: -ILI- the set of PWN synsets (labeled by their offsets in the database) taken as interlingual concepts: (077666677-n, 02564241 – v, 00933364-a, 00087007-b), the consortium selected a common set of ILI codes to be implemented for all languages, this selection took place in three steps: BSC1 (essentially the BC set of EWN): 1 218 concepts, BSC2: 3 471 concepts, BSC3: 3 827 concepts. Selection criteria for BCS1,2,3…(8 516 ILI-codes) were: number of languages in EWN linked to an ILI code, conceptual density, language specific criteria: each team proposed a set of concepts of interest and the maximum intersection set among these proposal became imperative

Serbian Wordnet is developed from the base concepts of the English WN using existing English/Serbian dictionaries in paper form, like a part of the Balkan wordnet project (BWN). Scientific teams which build Serbian Wordnet is Natural Language Processing Group on Faculty of Mathematics, University of Belgrade and they also added new specific Serbian and Balkan specific concepts and work on building SWN will be continued.

In order to describe the structure of SWN, we describe present statistical data and parameters about Serbian Wordnet. One of this parameter is the average number of literals per synset for SWN is 1.68. This ratio, however, significantly differs for different PoS (Part of Speech). Table 1 shows the PoS related distribution of synsets and literals (l), the literal/synset (l/s) ratio, the number of synsets with only one, two or three literals, and the maximum number of literals per synset (max). [3]

Table 1. Distribution of literals per synsets

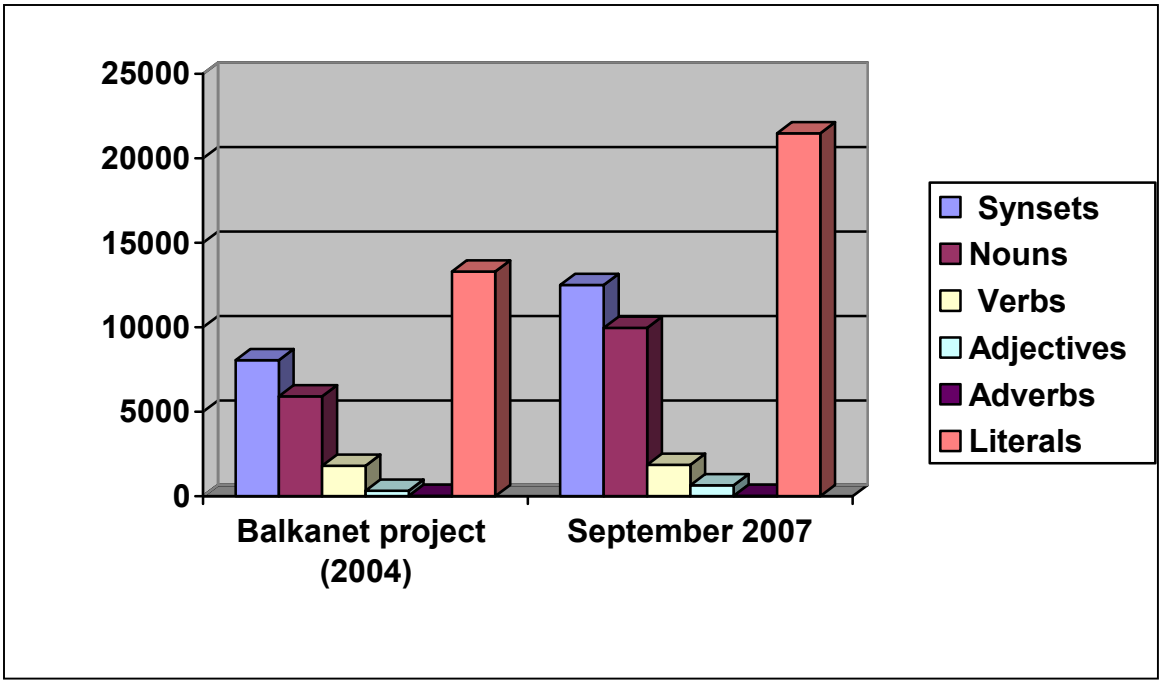
Part of speech (PoS)	synsets	percentage	**literals**	**l//s**	**1**	**2**	**3**	**max**
nouns	9964	79.81%	17164	1.72	4823	3710	1101	10
verbs	1864	14.93%	3487	1.87	778	700	287	11
adjectives	639	5.12%	807	1.26	501	113	22	8
adverbs	18	0.14%	18	1.00	1			1
total	12485	100%	21476	1.72	6120	4523	1320	



We also presents some statistical data about Serbian Wordnet for different Part of Speech- PoS (nouns, verbs, adjectives, adverbs) at the end of Balkan project and three years later, in September 2007 [4].

Table 2. Balkanet project - Final report about Status of the Serbian Wordnet

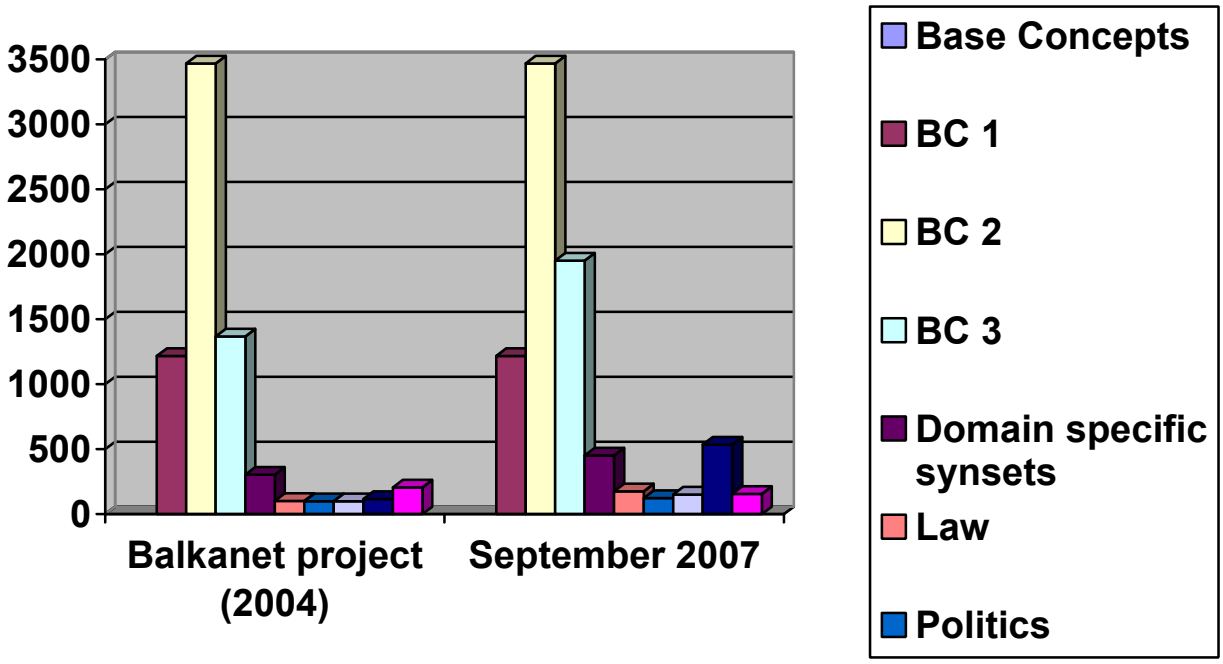
	Balkanet project - Final report (2004)	**September 2007**
Synsets	8059	12485
Nouns	5919	9964
Verbs	1803	1864
Adjectives	324	639
Adverbs	13	18
Literals	13295	21476
The medium length of synsets	~ 1.625	1.72
The average number of senses per literal	~ 1.21	1.14



We explained structure of Base concepts, but there are also different kind of synsets according to domain. They are shown below obtained results in Table 3. Very interesting synsets are BalkaNet and Serbian Specific-Concepts [5]. The explanation could be that there exist significant historical, social and cultural links between the Balkan languages represented in the BalkaNet project. All languages in the project have significantly influenced each other and there are several concepts that are specific to this region of the world. With these considerations in mind, the BalkaNet consortium made an attempt to incorporate these shared concepts in the wordnets of the respective languages in a systematic way.

Table 3. Balkanet project - Final report about Status of the Serbian Wordnet

	Balkanet project - Final report (2004)	**September 2007**
Base Concepts		
BC 1	1219	1219
BC 2	3469	3469
BC 3	1369	1953
Domain specific synsets	305	453
Law	103	176
Politics	101	125
Economy	101	152
Balkan specific synsets	117	535
Serbian specific synsets	206	156



The development of Serbian Wordnet continues after Balkanet with its enhancement in various domains. One of the domains especially well and systematical covered is the domain of biomedical disciplines which includes microbiology, zoology of Invertebrates and Vertebrates, cytology, embryology, histology, veterinary, agriculture and others. Biomedical disciplines are numerous and very dynamically develop. Almost every day we have new discoveries in scientific literature. Many scientific terms from biological and medical fields, scientists and researcher do not translate in Serbian and use them during their work and in literature.

We used various traditional and Internet resources for the enhancement of SWN in domain of biomedicine and show possible practical usage of SWN in information retrieval of this domain. Only in one year, 2006. we added about 500 aligned synsets for scientific field – biomedicine or more precisely 462. We added new concept for six ontological categories in Serbian Wordnet, based on SUMO ontology: Genetics, Virus, Bacterium, Cell, Science Fields and Microorganism. Nowadays, Princeton Wordnet consisting of approximately 140,000 synsets and Serbian Wordnet of 13 000 synsets. In the last two decades, Wordnet has evolved as the most comprehensive computational lexicon of general English and many other languages from all over the world. The next phase will be creation of an entirely new kind of information resource for public health, Medical Wordnet, integration of Wordnet and one of the oldest and the best medical database – Medline[6].

We can show one practical example, synset *genetic engineering*, in View (Figure 3.), Tree (Figure 4.) and XML format (Figure 5.) in VisDic.

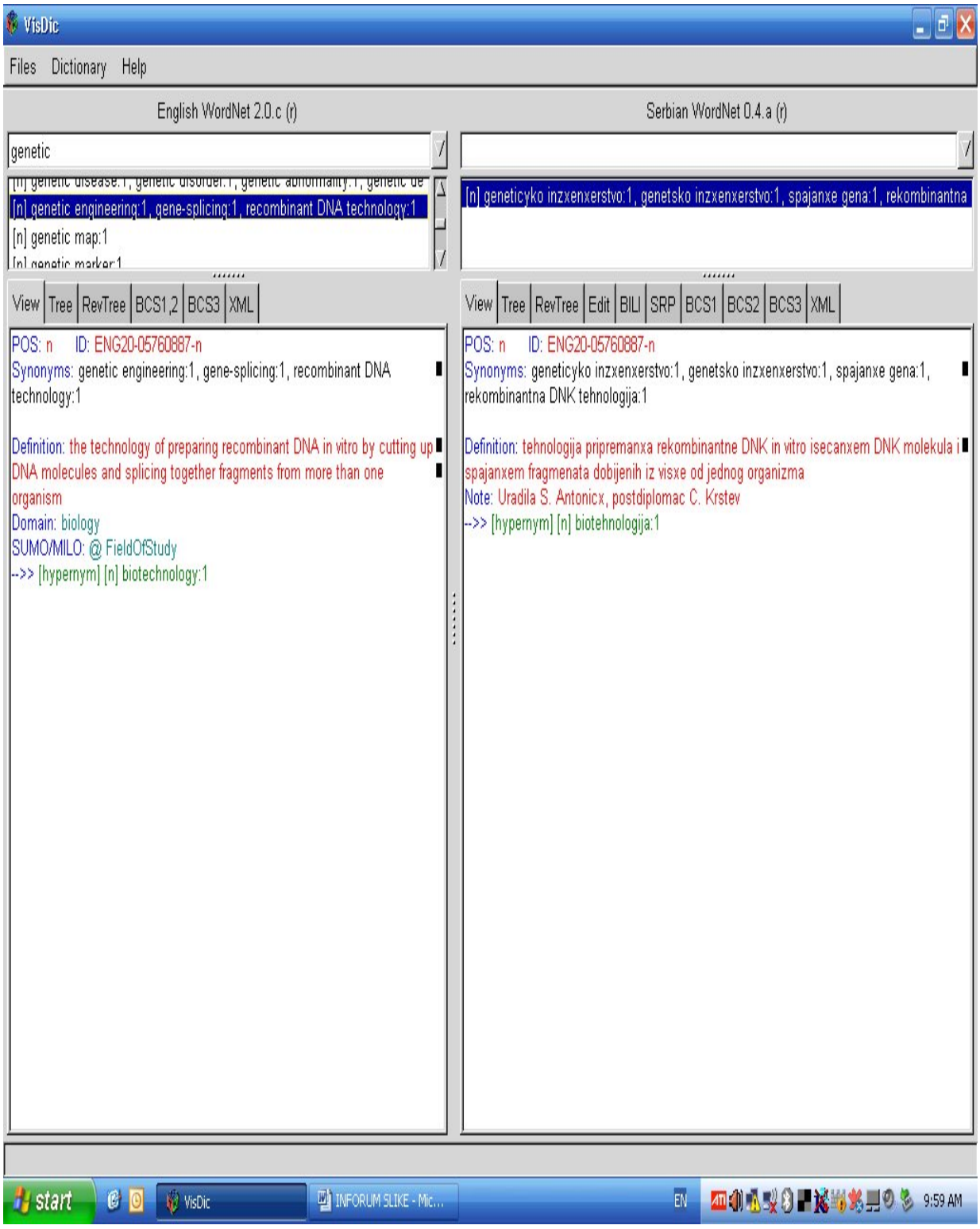


Figure 3. View format

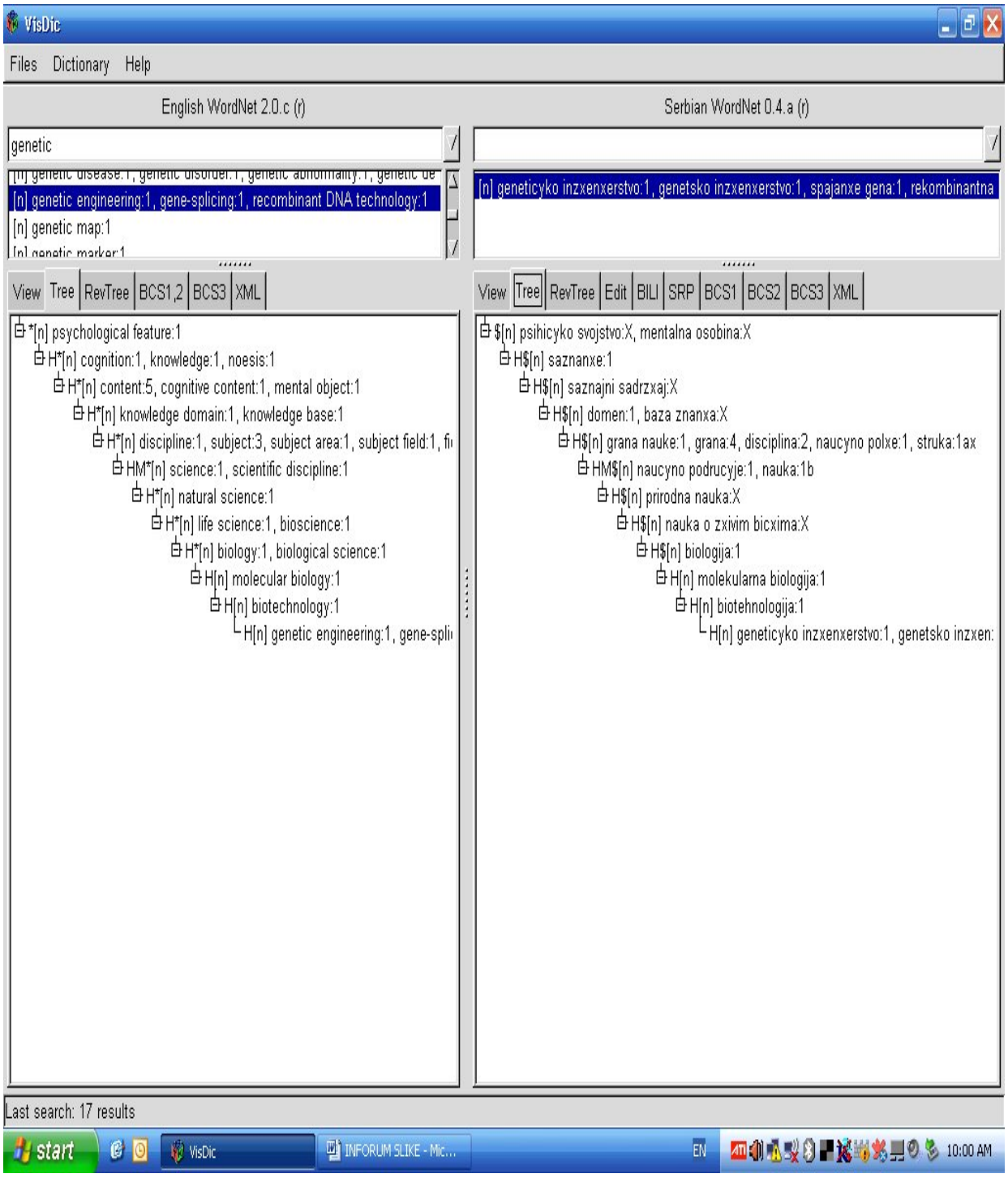


Figure 4. Tree format

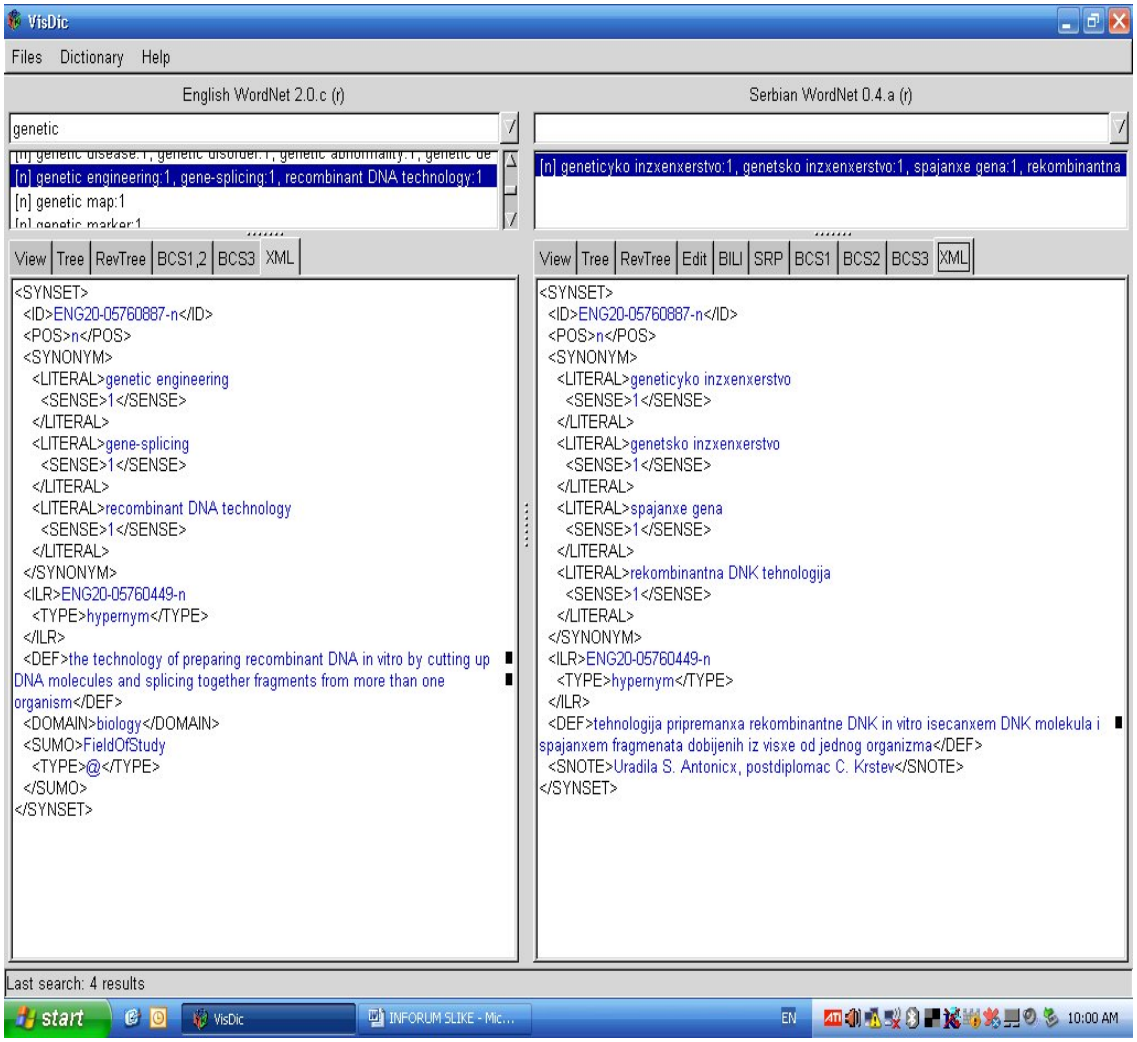


Figure 5. XML format

Last but not least a preliminary approach of Wordnet, including Serbian Wordnet is consideration that semantic networks are valuable in the context of real world systems and user communities. The ultimate objective of this contribution is to spread the knowledge and experience that we have acquired, to the benefit of the research and industrial communities.

Piek Vossen, leader of EuroWordNet project in 2004. wrote that Slavic languages are a bit more alien to English than the other West-European languages. Both morphological structures and cultural background are slightly different. Borders are not only removed physically and politically but, perhaps more importantly, by enabling communication. The BalkaNet project has connected the East with the West of Europe in a very special and unique way [7]

.

References:

[1] Miller G.A., Beckwidth R., Fellbaum C., Gross D., Miller K.J. (1990). Introduction to WordNet: An On-Line Lexical Database. In *International Journal of Lexicography*, vol. 3, no. 4), pp. 235-244.

[2] Tufiş D, Cristea D., Stamou S. (2004) BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. Romanian Journal of Information Science and Technology Special Issue, vol. 7, No. 1-2, p. 9-43

[3] Krstev C., Pavlović-Lažetić G., Vitas D., Obradović I. (2004) "Using Textual and Lexical Resources in Developing Serbian Wordnet", in Romanian Journal of Information Science and Technology", vol. 7, No. 1-2, pp. 147-161, Romanian Academy, Publishing House of the Romanian Academy.

[4] Balkanet final report

http://www.ceid.upatras.gr/Balkanet/deliverables/finalreport_sub.pdf

[5] Krstev, C., Vitas, D., Stanković, R., Obradović, I., and Pavlović-Lažetić, G. (2004). Combining Heterogeneous Lexical Resources. Proceedings of the 4th International Conference on Language Resources and Evaluation LREC 2004, Lisbon, Portugal, 1103-1108.

[6] Fellbaum C., Hahn U., Smith B. (2006) Towards new information resources for public health—From WORDNET to MEDICALWORDNET Journal of Biomedical Informatics 39, 321–332.

[7] Vossen, P. (2004) Introduction to the Special Issue on the BalkaNet Project. Romanian Journal of Information Science and Technology Special Issue, vol. 7, No. 1-2, p. 5-6.