COOPERATIVE WORK IN FURTHER DEVELOPMENT OF SERBIAN WORDNET

Cvetana Krstev, Faculty of Philology, Belgrade, **Bojana Đorđević**, Faculty of Philology, Belgrade,
Sanja Antonić, University Library "Svetozar Marković", Belgrade,
Nevena Ivković-Berček, Faculty of Theology, Belgrade,
Zorica Zorica, Art Gallery, the Legacy of the Collector Rajko Mamuzić, Novi Sad,
Vesna Crnogorac, Association of the Librarians of Serbia, **Ljiljana Macura**, National library of Serbia

Abstract: In this paper we present the computational lexical database Wordnet that has become the *de facto* standard for semantic networks. The first network of this type and, at the same time, the most developed one that was produced at the University of Princeton in the Laboratory for cognitive sciences, has served as a basis for the development of wordnets for many other languages. Wordnets that were developed in the scope of European research projects EuroWordNet and BalkaNet are aligned with the Princeton wordnet, which enables their successful use in many multi-lingual applications. The development of the Serbian wordnet was initiated during the BalkaNet project, and, after the end of that project, the work on its further development continued through the volunteer and cooperative work of many collaborators whose work is presented in this paper.

1 Wordnet

Wordnet is an extremely large lexical database that is organized through nodes and relations established between these nodes (Fellbaum, Christiane, ed. 1998). These nodes, which are termed in Wordnet *synsets*, from *synonymous sets*, represent sets of words that express in some context the same meaning. For instance, in English wordnet one synset is {teacher:1, instructor:1} with the meaning "a person whose occupation is teaching". Numbers that follow the words in this set (in the given example in both cases it is the number "1") indicate that with these words the given meaning can be expressed, but that in a different context some other meaning can be expressed as well. Really, the synset {teacher:2} corresponds to the meaning "a personified abstraction that teaches", as in the example "experience is a demanding teacher". This example shows that although the words *teacher* and *instructor* are synonyms in the first context, they are not in the second, since in the last example the word *teacher* cannot be replaced by the word *instructor* (*experience is a demanding instructor).

This database is divided, according to part-of speech, into nouns, verbs, adjectives, and adverbs. The nominal part of the database is organized as a hierarchy of nodes, which is established on the basis of the relation of subordination and superordination between the meanings represented by corresponding nodes. One notion is subordinated to another notion not only if it has all the features of the superordinated notion, but also if it has some specific additional features as well. This can be examplified by the part of the hierarchy to which the noun's synset {teacher:1, instructor:1} belongs. Its direct superordinated synset is {educator:1, pedagogue:1} (someone who educates young people), whose direct superordinated notion is {professional:1, professional person:1} (a person engaged in one of the learned professions), whose direct superordinated notion is {adult:1, grownup:1} (a fully developed person from maturity onward), etc. This example shows that each superordinated notion looses some of its features, and, in that respect, represents a more abstract notion than its subordinates. Moreover, almost every superordinate notion has more than one subordinate notion. For instance, a grownup can be engaged in a learned profession, or he/she can be able to do a variety of different jobs acceptably well ({Jack of all trades:1}; also, a professional can be engaged in the education

of young people or he/she can be authorized to practice law, conduct lawsuits or give legal advice ({lawyer:1, attorney:1}) etc.

The relation of subordination and superordination (or "hypernym-hyponym" relation) is not the only one that can be established between meanings. The exhaustive analysis of notions that are lexicalized by nouns and various relations existing between them is given in (Miller, George A., 1990). Some of these relations are implemented in Princeton wordnet (wordnet.princeton.edu) thus establishing a complex network of lexicalized concepts. Relations that are prevailing in the Princeton wordnet, apart from subordination and superordination, are relations "a part ↔ a whole" and "a member ↔ a whole". As an example, we can examine the concept represented by the synset {warship:1, war vessel:1, combat ship:1} (a government ship that is available for waging war), that is "a member" of a *fleet*, a concept that is in the Princeton wordnet represented by the synset {fleet:4} (a group of warships organized as a tactical unit). On the other hand, warship, contains as its part a *naval gun*, that is, {naval gun:1} that represents a kind of a naval weaponry carried on a warship. Similar with these is a relation "a component ↔ a whole" which connects concepts, such as {protein:1} (any of a large group of nitrogenous organic compounds that are essential constituents of living cells) and {egg:1} (oval reproductive body of a fowl used as food).

Another important relation that is established between noun synsets is antonymy that connects the concepts that have (almost) opposite meaning. Obvious examples are {female:2, female person:1} (a person who belongs to the sex that can have babies) and {male:2, male person:1} (a person who belongs to the sex that cannot have babies) and {sorrow:1} (an emotion of great sadness associated with loss or bereavement) and {joy:1, joyousness:1, joyfulness:1} (the emotion of great happiness).

Another important relation that is established between the synsets in the Princeton wordnet is the relation that connects the concepts that are lexicalized by different parts of speech. The important relation that connects a noun synset with a adjective synset is the relation "be in state ↔ the state of". One example is the synset {cleanness:1} (the state of being clean; without dirt or other impurities) that is connected with the adjective synset {clean:1} (free from dirt or impurities; or having clean habits). The relation of antonymy is frequent among the adjectivs; so, the synset {clean:1} is connected by the relation "near antonym" with the synset {dirty:1, soiled:1, unclean:1} (soiled or likely to soil with dirt or grime). This synset is again in relation to the noun synset {dirtiness:1, uncleaness:1} (the state of being unsanitary) through the relation "be in state ↔ the state of", while this synset is in its turn again connected by the relation "near antonym" with the initial synset {cleanness:1}. If we add to all this the relations that are established between verbal synsets, such as "causes ↔ caused by" that connect, for instance, synsets {stand:10; stand up:2; place upright:1} (put into an upright position) and {stand:1; stand up:4} (be standing; be upright) it is clear that the Princeton wordnet represents a dense network of nodes and relations.

The nature of these relations is different. Some of them are symmetric, such as "near antonym", because if *A* has an (almost) opposite meaning of *B*, then *B* has an (almost) opposite meaning of *A*. Some of the other relations are asymmetric, as a relation "causes", because if *A* causes *B*, then *B* does not cause *A*. The asymmetric relations, however, always have a counter relation. For instance, in regards to the relation "causes" its counter relation is "caused by". Thus, if *A* causes *B*, then *B* is therefore caused by *A*. Some of the relations are by their nature "one-one", such as "near antonym", while the others are usually "many-one", like the relation "hypernym-hyphonym". Namely, usually for one *A* exists at most one *B* that has an (almost) opposite meaning, and vice versa. On the other hand, while an *A* usually has a unique superordinated concept *B,* in most

cases, *B* in its turn has more than one subordinated concept. The relation "a part ↔ a whole" is by its nature a "many-many" relation because an *A* can be a part of various *B,* while the same *B* can have, besides *A,* many other components. For instance, {handle:1, grip:2, handgrip:1, hold:8} (the appendage to an object that is designed to be held in order to use or move it) is a part of many different objects: brush, suitcase, fry pan, umbrella, to mention just a few, and it is clear that all of these objects contain a handle besides many other components. This discussion about the nature of relations established between the concepts in a wordnet is important from the standpoint of the implementation of the lexical database itself. In order to avoid redundancy, for asymmetric relations, only one relation is recorded (and not its counter relation), and that is the relation that in most cases has unique value. For instance, a database records the usually unique hypernym *B* of the concept *A*; the potentially numerous hyponyms of the concept *B* are not recorded since they can be indirectly derived form the recorded counter relation.

The Princeton wordnet has had a great impact not only because it represents a vast database but also because it has been applied in many different fields, such as automatic sense disambiguation, term expansion in information retrieval, and construction of structured representations of document content. In actuality, it has become so popular that it can almost be considered a *de facto* standard in natural language processing. The many uses of the Princeton wordnet are presented in (Fellbaum, Christiane, ed. 1998). Princeton wordnet is continually updated, and the last version 2.1 for the operating system Windows was released in March 2005.

2 The Enhancements of the Princeton Wordnet

The structure of the Princeton wordnet database was enhanced several times with additional information in order to make it even more usable in various natural language applications. We will here present two of these enhancements that were significant for the project that is the main topic of this paper.

The first enhancement added semantic domains to the Princeton wordnet. Semantic domains provide a natural way to establish semantic relations between the meanings of words that can be useful in many natural language processing applications. Semantic domains are areas of human interests such as *sports*, *economics* or *politics*, which exhibit their own terminology and lexical coherence. The usage of domains is known to linguistics (for the description of semantic fields), as well as to lexicography (for subject field codes).

The Princeton wordnet was augmented by adding at least one domain label to almost every synset. Domain labels were chosen from approximately two hundred hierarchically organized domains. For instance, synset {mouse:1} (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails) belongs to the domain *zoology*, while synset {mouse:2; computer mouse:1} (a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad) belongs to the domain *computer science*. This new information on domain supplements the existing information in the wordnet. One domain can include synsets that belong to various parts of speech and to different hierarchies. The additional benefit is that domains may group meanings of the same word into homogeneous clusters, thus reducing word polysemy, which is in the wordnet very large since the meanings are very finely separated. Let us take as an example the noun *time* that occurs as a simple word (that is, not as a component of a compound) in eight synsets of the Princeton wordnet of which five belong to the domain *time period*.

Two hundred domain labels from the Dewey Decimal Classification, hierarchically organized

in a tree, were used for synset labeling. At the top of this domain hierarchy are domains: *doctrines*, *free_time*, *applied_science, pure_science*, *social_science* and *factotum*, where the label *factotum* was used in cases where no other domain could be applied. On the top are also the domains *number*, *color*, *time-period*, *person* are *qualities* that are not further refined. On the other hand domain *doctrines*, has as its sub-domains *archaeology*, *astrology*, *history, linguistics*, *literature*, *philosophy, psychology*, *art* and *religion*, most of which are further refined. More about the project of Princeton wordnet enhancement with semantic domains can be find in the articles "Integrating Subject Field Codes into WordNet" (Magnini, B. and Cavaglià, G. 2000), "Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing" (Bentivogli, L. et. al. 2004) and at the web site wndomains.itc.it.

The second enhancement of the Princeton wordnet is related to its linking-up with the SUMO ontology (Standard Upper Merged Ontology) whose development as a new standard initiated IEEE in 2000. By ontology in this context a dictionary or a glossary is considered that has a structure, which enables the computer analysis of its content. One such ontology consists of concepts, axioms and relations that describe a domain of interest. An *Upper Ontology* is limited to meta-concepts that are abstract and generic in its nature and therefore general enough to cover a wide range of domains in the upper level. The concepts that are specific to some particular domains are not included in the upper ontology. The term 'Merged' in the name of the ontology indicates that it was developed by merging publicly available ontological content into a single comprehensive and coherent structure (Pease, A., Niles, I., 2002).

So that it could be easier used and applied, the SUMO consists of a relatively small number of assertions and rules: approximately 4,000 assertions (including over 800 rules) and 1,000 concepts. Some of the general topics covered in the SUMO include:

- Structural concepts such as instance and subclass
- General types of objects and processes
- Abstractions including set theory, attributes, and relations
- Numbers and measures
- Temporal concepts, such as duration
- Parts and wholes
- Basic semiotic relations
- Agency and intentionality.

The question is how ontology can be successfully used in various natural language applications that process free texts, such as texts that were not preprocessed and whose structure was not made explicit. One answer to this question offers the connection of the SUMO with an extensive lexical database like the Princeton wordnet (Niles and Pease 2003). The linking had been established with the version 1.6 of Princeton wordnet, but the established links were transferred to all later versions of this wordnet. More about SUMO, as well as the browsing of its hierarchy, is provided on the sites sumo.ieee.org and www.ontologyportal.org.

Having in mind the fact that SUMO consists of a relatively small number of concepts, while, wordnet represents a rich lexical database that consistsed at the moment of the linkage of almost 100,000 synsets it is necessary to specify how the connection between an SUMO concept and one wordnet synset was established. Basically, three types of relations were used: synonymy, hypernymy, and instantiation. These relations will be illustrated by a few examples. In Princeton wordnet exists the synset {battle:1, conflict:3, fight:4, engagement:1} (a hostile meeting of opposing military forces in the course of a war) which is synonymous with the concept "Battle" from the SUMO; thus, the information "= Battle" is added to the synset. This synset has as one of its hyponyms the synset {naval battle:1} (a pitched battle between naval fleets) for which, natu-

rally, the synonymous concept does not exist in the SUMO. In such cases the synset is linked to its superordinated concept, which in our example means that to the synset {naval battle:1} the information "+ Battle" is added. Finally, the synset {Iwo:1, Iwo Jima:2, invasion of Iwo:1} (a bloody and prolonged operation on the island of Iwo Jima in which American marines landed and defeated Japanese defenders (February and March 1945)) represents one instance, or a case of a battle. To such synsets the third kind of relation is applied that indicates that the concept represented by the wordnet represents one member of the class denoted by the SUMO concept. In such a case the label "@ Battle" is added to the synset. It is not surprising that there are cases when more than one synset in the wordnet was connected by the relation of synonymy with the same concept from the SUMO. The difference between synsets can be linguistically important but from the standpoint of knowledge engineering, quite irrelevant. The information "= Battle" was thus associated not only to the synset {battle:1, conflict:3, fight:4, engagement:1} but also to the synsets {invasion:1} (the act of invading; the act of an army that invades for conquest or plunder) and {combat:1, armed combat:1} (an engagement fought between two military forces).

In order to elaborate this example further, we present the branch of the hierarchy to which the ontological concept "Battle" belongs:

entity → physical → process → intentional process → social interaction → contest → violent contest → battle

Top of the hierarchy tree of sub-classes looks like this:

entity→
- physical→
  - object→
  - process→
- abstract→
  - quantity→
  - attribute→
  - set or class→
  - relation→
  - proposition→
  - graph→
  - graph element→

Therein, the described mapping of the Princeton wordnet into the SUMO can function as a natural language index for the concepts from ontology, as well as a bridge between the structured concepts from the SUMO and free texts that are subjected to processing in the scope of various applications, such as conceptual indexing (Stamou, S. et al. 2004) and text classification (Tufiş, D. and Koeva S. 2007).

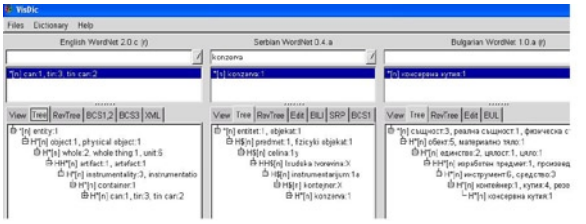
3 Wordnet and Multilinguality

As a resource that promises very much in natural language applications, the Princeton wordnet has become very popular and has initiated many projects whose aim was the production of similar resources for other languages. One of the first projects in this line was the project EuroWordNet that had been funded by the European community in the framework of the FW4 programme from 1996 to 1999. The aim of this project was the development of a multilingual lexical database that would contain wordnets for eight European languages: English, Dutch, Italian, Spanish, French, German, Czech, and Estonian. The structures of all these wordnets corresponded to the structure of the Princeton wordnet, but each wordnet was independent and contained concepts that corresponded to the specific lexicalization in each particular language. Yet, in order to satisfy the needs of multilingual applications that are becoming more and more important in Internet and web context, the project EuroWordNet introduced the notion of an *Inter-Lingual-Index* (abbr. ILI) through which a synset in one language is connected to the synsets in other languages that represent similar concepts (Vossen, P. 2004). The purpose of the Inter-Lingual-Index is to provide an efficient mapping across the autonomous wordnet structures of individual languages. Since each monolingual wordnet is a separate ontology, ILI itself is reduced to a condensed and universal index of meaning. Keeping in mind the independence of individual wordnets and the assumed differences in lexicalization of

the concepts in different languages, synsets are connected via ILI through different relations that are by their nature "many-to-many". Multilingual database thus conceived offers, besides the connection of synsets from one language with related synsets in other languages, the possibility to share knowledge that is language independent. ILI, thus, enables the usage of additional knowledge that was introduced into the Princeton wordnet through domain labels and SUMO concepts, which was discussed in the previous section, by other wordnets of other languages that are connected with it.

Along the same lines, the BalkaNet project, funded by the European Commission from 2001 to 2004, set as its goal the development of aligned semantic networks for Bulgarian, Greek, Romanian, Serbian, and Turkish, while at the same time extending the existing network for Czech, initially developed within the EuroWordNet (Tufiş, D. et al. 2004). Six teams were formed, each responsible for the development of a wordnet in one of the six languages. The main aim of the BalkaNet project was the development of a modern language resource for Balkan languages that would enable new access to information that is expressed within Balkan languages. In addition, the aim was to enhance the multilingual database that was established within EuroWordNet with Balkan languages.

The main activity of the BalkaNet project was the development of individual wordnets for Balkan languages and their connection with the existing lexical database EuroWordNet. These main activities were planned and realized synchronously, which means that the core of each monolingual wordnet was built from several commonly agreed sets selected by PWN. Beyond these sets the network for each language has been developed independently, but always within the framework set by the Princeton wordnet. This approach had generated some specific problems. Namely, during the work on the development of the network the following questions had often been raised: are concepts linguistically independent or not, are the lexicalization patterns for concepts universal, is the structure of PWN valid for other languages as well, is the set of semantic relations built in PWN sufficient for all languages (Vossen, P. 2004). Although the work on the development of specific networks for Balkan languages often pointed to a negative answer to these questions, the initially established procedure has not been abandoned. As wordnet type networks are being developed today, mainly for information science purposes, the main application of these networks is foreseen in their incorporation into information science applications based on natural language processing, such as a network-based classification of documents and multilingual search, where the existence of a multilingual database with mutually aligned concepts is crucial.



The figure represents the lexicalization of the notion "airtight sealed metal container for food or drink or paint etc." in English ({can:1, tin:3, tin can:2}), Serbian ({konzerva:1}) and Bulgarian ({консервна кутия:1}). While the lexicalization of this notion is not an issue, its hypernyms that are in English lexicalized as {container:1}, {instrumentality:1, instrumentation:1} and {artifact:1, artefact:1} are rather artificially mapped from the Princeton wordnet to both Serbian and Bulgarian wordnets.

4. Development of the Serbian Wordnet

Development of the Serbian wordnet started within the BalkaNet project. By the end of the project the Serbian database consisted of 8,059 synsets, 7,736 of which were adopted from the Princeton wordnet, while 117 belonged to the set of the so-called Balkan specific concepts, and 206 belonged to the set of Serbian specific concepts. One concepts specific to the Balkans

that is well-known and lexicalized in all Balkan languages and which does not exist in the Princeton wordnet is {alva:1} in Serbian, {халва:1} in Bulgarian, {χαλβάς:1} in Greek, {halva:1} in Romanian and {kağıt helva:1} in Turkish. The work on the development of the Serbian wordnet did not stop with the formal end of the BalkaNet project; however, it did not proceed in the frame of a formal project, but rather relied on the contribution of volunteers. In the first year after the end of the BalkaNet project, development was oriented towards the synsets from those domains of biology that deal with plant and animal species, as well as higher classification groups to which these species belong. The choice of domains was synchronized with the enhancement of Serbian electronic dictionaries with entries from the same domains. During this period, considerable work was done on the extension of the Serbian wordnet with Balkan specific and Serbian specific concepts (Krstev, C. 2006).

Cooperative work on further development of the Serbian wordnet started in early 2006. This was done mainly by many postgraduate students (namely those who had beforehand obtained a graduate diploma in various fields and are presently employed as librarians, informaticians or documentalists in public or special libraries) at the Group for Library and Information Sciences at the Faculty of Philology of the University of Belgrade. As one task of the obligatory subject System of Scientific Information in these postgraduate studies, students have to produce a seminar work. The idea has emerged, having in mind their present occupation that these students might develop one segment of the Serbian wordnet using the specific knowledge that they have obtained during their previous education. Subsequently the volunteers from other postgraduate groups were recruited, and all of them showed a great enthusiasm for this commonly held work and this work will be presented in the following section.

We have used the software tool WS4LR (*Work Station for Lexical Resources*) that is described in more detail in (Obradović, I. and Stanković, R. 2008) to select the subsets of synsets that best suit the interests and qualifications of the students involved. Since both Serbian and Princeton wordnets use the common XML format introduced in the BalkaNet project for the transfer and exchange, this tool enables the user to formulate their own *XML Path* expressions by which he/she can select a subset of synsets from the chosen wordnet. For such a selection the combined information on domain and ontological category was usually used, as, in most cases, subsets obtained by the use of the information on domain only were too large. For instance, the Princeton wordnet has 1,181 synsets that belong to the domain of law, and our view was that the preparation of such a large number of synsets would be too demanding a task for a seminar work of just one student. On the other hand, the usage of only the ontological categories could put into the same subset notions that belong to various domains, including those for which the students do not have adequate competence. For instance, the ontological category "Charcater" belongs not only to the domain "linguistics", for which we were interested, but also to the domains "factotum", "number", "publishing", etc. Also, in order to exhaustively cover the chosen domain, it was sometimes necessary to make amendments to the selected subset. For instance, one selected subset from the domain of linguistics referred to the ontological category "NaturalLanguage". Processing an inclusion of this very large subset to the Serbian wordnet showed that still some of the "great" European languages were not included, like Russian, French, etc. It was revealed that these languages were connected to the separate ontological categories "=RussianLanguage" and "=FrenchLanguage". These missing synsets were subsequently included in the Serbian wordnet.

4.1. The Domain of Linguistics

Bojana Đorđević

I joined the project for the development of the Serbian wordnet in early 2007, after I had graduated from the Department of General Linguistics

at the Faculty of Philology in Belgrade. My task in this project was to analyze synsets from the domain of linguistics of the Princeton wordnet and their adaptation for the Serbian wordnet. The processed set of synsets from the domain of linguistics consisted of the following ontological categories: morphemes (16 synsets), grammars (238 synsets), characters (87 synsets) and natural languages (595 synsets). Keeping in mind that some synsets were added afterwards in order to secure the correct hierarchical connection of this subset with the remaining Serbian wordnet, from the domain of linguistics a total of 946 synsets were processed.

A typical example of one translated synset that is connected to the ontological category "NaturalLanguage" is given below:

<SYNSET> <ID>ENG20-06480396-n</ID> <POS>n</POS> <SYNONYM> <LITERAL>mother tongue <SENSE>1</SENSE> </LITERAL> <LITERAL>maternal language <SENSE>1</SENSE> </LITERAL> <LITERAL>first language <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR> <TYPE>hypernym</TYPE> ENG20-06479855-n </ILR> <DEF>one's native language; the language learned by children and passed from one generation to the next</DEF> <DOMAIN>linguistics</DOMAIN> <SUMO>NaturalLanguage <TYPE>+</TYPE></SUMO> </SYNSET>	<SYNSET> <ID>ENG20-06480396-n</ID> <POS>n</POS> <SYNONYM> <LITERAL>maternji jezik <SENSE>1</SENSE> </LITERAL> <LITERAL>prvi jezik <SENSE>1</SENSE> </LITERAL> <LITERAL>rođeni jezik <SENSE>1</SENSE> </LITERAL>[1] </SYNONYM> <ILR> <TYPE>hypernym</TYPE> ENG20-06479855-n </ILR> <DEF>jezik koji je najpre usvojen u detinjstvu ili onaj kome se daje prednost u višejezičnoj situaciji</DEF> <SNOTE>Uradila B. Đorđević, postdiplomac C. Krstev</SNOTE> </SYNSET>

During my work on the Serbian wordnet I was confronted with several kinds of problems. Since many synsets dealt with the names of languages that had to be included in the Serbian wordnet, the main problem was to find an adequate and, if possible, adopted name for some less investigated languages that were represented in the Princeton wordnet in detail. This was especially the case for many Amerindian languages, but also for some groups of Afro-Asiatic languages.

Another kind of problem occurred in the attempt to find an adequate term for the phenomenon that either does not exist in Serbian or is classified in a different way. The following example is related to the kind of the grammatical object that does not exist in Serbian. The author of the used term is Ljiljana Mihailović (Mihailović, Lj.1967):

<SYNSET> <ID>ENG20-05923070-n</ID> <POS>n</POS> <SYNONYM> <LITERAL>retained object <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR> <TYPE>hypernym</TYPE>ENG20-05922459-n </ILR> <DEF>an object in a passive construction</DEF> <DOMAIN>grammar</DOMAIN> <SUMO>NounPhrase <TYPE>+</TYPE> </SUMO> </SYNSET>	<SYNSET> <ID>ENG20-05923070-n</ID> <POS>n</POS> <SYNONYM> <LITERAL>zadržan objekat <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR> <TYPE>hypernym</TYPE> ENG20-05922459-n </ILR> <DEF>objekat u pasivnoj konstrukciji</DEF> <SNOTE>Uradila B. Đorđević, postdiplomac C. Krstev</SNOTE> </SYNSET>

Similarly, some concepts that in English can be expressed by a simple word (*to punctuate*) had to be expressed in Serbian descriptively (*obeležiti znacima interpunkcije* 'to mark by punctuation marks').

The greatest number of the terms and glosses were taken from the Serbian translations of the Cambridge Encyclopaedia of Languages (Kristal, 1995) and the Encyclopaedic Dictionary of Modern Linguistics (Kristal, 1998), and *Uvoda u opštu lingvistiku* 'An Introduction to General Linguistics' (Bugarski, 1996) and

Gramatike engleskog jezika 'Grammar of the English Language' (Đorđević, 2002). The studies *Jezici* 'Languages' (Bugarski, 1996a) and *Jezik* 'Language' (Sapir, 1992) were consulted for translating the names of these languages, and they were especially useful for the translation of names of the Amerindian languages. Many available general dictionaries were also used as needed: an "English Language Dictionary" (Institut za strane jezike, 2005), a "Dictionary of Foreign Words" (Klajn, Šipka, 2006) and a "Dictionary of Synonyms" (Ćosić, 2007).

The Internet databases were practically invaluable for the translation of the names of these languages. One of the main sources was Wikipedia, both the Serbain and Croatian versions. I have also used the list of basic language groups and sub-groups compiled by Danka Šipke, that represents a part of the wider thematic list devoted to students that study Serbian as a foreign language. I should also mention the Croatian database *PhraseBASE* which groups together large numbers of languages in language families, as well as the text by Nevenka Hajdarović *Izmjene i dopune u UDK i njegova primjena u BH Bibliotekarstvu* 'the Amendments of UDK and its application in libararianship in Bosnia and Hercegovina', which contains a comprehensive list of world languages. I have also used the rich base of the names of languages to which the automatic translator PROZ translates for Serbian and vice versa.

When the comparison of resources from the Internet yielded an unrealistic number of possible translations, the frequencies of their occurrences obtained by Google had the final say in the decision process. However, in some situations the frequencies and occurrence or non-occurrence of some terms in Google results could not be decisive, as linguistic texts in Serbian are still scarce.

On several occasions I have obtained useful information from members of the mailing list *ST-L* that had been initiated by Prof. Dr Danko Šipka as an initial step in planning the development of one Serbian electronic corpus. The list was initially dedicated to the discussion about the content of a corpus of Serbian, as well as Croatian and Bosnian, and the form of texts that would enter it. Later however the discussions turned to linguistic topics. I am especially thankful for the general language advice that I obtained from Wales Brown, the professor of linguistics at Cornell University and Pavle Ćosić, the linguist and author of the dictionary of synonyms.

4.2. The Domain of Biomedicine

Sanja Antonić

Practical work on the development of the Serbian wordnet represents the last chapter of my master thesis "The Development of a Computational Semantic Network for Biomedical Sciences". The mentor for this thesis was Prof. Cvetana Krstev. I work at the University Library "Svetozar Marković" in Belgrade, in the Department for Scientific Information as a senior research assistant for biomedicine and biotechnology. I have studied molecular biology and physiology and have graduated at the Faculty of Biology, University of Belgrade. The knowledge that I have acquired during my studies, I have enhanced and applied throughout the course of my work on the domain of biomedicine for the Serbian wordnet.

Biomedicine encompasses many scientific disciplines, most of which are very dynamic and develop quickly. Theory and terminology in biomedicine are becoming obsolete very quickly, so it is necessary to keep in touch with new knowledge and discoveries on a daily basis. As a result of this, many scientific and technical terms are often used in their original form and are not translated. During my work I had to keep in mind the audience to which the network is dedicated, and sometimes it was difficult to harmonize the needs of professionals for the processed fields with those of common users. Moreover, I had to maintain the usability of the network in computer applications.

My work on the development of the Serbian wordnet involved the following scientific disci-

plines, each having their own characteristics and terminology: cytology, histology, embryology, genetics, microbiology, zoology of invertebrates and vertebrates, veterinary medicine, agriculture, and so on. More precisely, I worked on the Serbian adaptation for those parts of the Princeton wordnet that belong to the domain of biology and, according to the SUMO, are connected to the following ontological categories: Cell, Genetics, Virus, Bacterium, Microorganism, ScienceFields. The total number of processed synsets was 462.

A typical example of a translated synset that is related with the ontological category "Microorganism" was processed in the following way:

<SYNSET> <ID>ENG20-01298897-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>mycoplasma <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-01280902-n <TYPE>hypernym </TYPE> </ILR> <ILR>ENG20-01298746-n <TYPE>holo_member </TYPE> </ILR> <DEF>the smallest self-reproducing prokaryote; lacks a cell wall and can survive without oxygen; can cause pneumonia and urinary tract infection</DEF> <DOMAIN>biology </DOMAIN> <SUMO>Bacterium <TYPE>+</TYPE> </SUMO> <RILR>ENG20-01299130-n <TYPE>hypernym </TYPE> </RILR> </SYNSET>	<SYNSET> <ID>ENG20-01298897-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>mikoplazma <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-01280902-n <TYPE>hypernym </TYPE> </ILR> <ILR>ENG20-01298746-n <TYPE>holo_member </TYPE> </ILR> <DEF>najmanje prokariote koje mogu da se bespolno razmnožavaju; nedostaje im ćelijski zid i mogu preživeti bez kiseonika; mogu izazvati pneumoniju (upalu pluća) i infekcije urinarnog trakta </DEF> <SNOTE>Uradila S. Antonić, postdiplomac C. Krstev</SNOTE> <RILR>ENG20-01299130-n <TYPE>hypernym </TYPE> </RILR> </SYNSET>

Mycoplasmas are microorganisms that are highly characteristic. The gloss given in the Princeton wordnet was "the smallest self-reproducing prokaryote". However, the knowledge about the life cycle of organisms shows that a more appropriate definition would be "the smallest prokaryote with asexual reproduction" and the gloss for the synset in the Serbian wordnet has been changed according to that. This example shows that since our aim was to produce a high quality wordnet, not even the glosses were acquired automatically.

The synset {paramecium:1, paramecia:1} represents organisms from the genus Paramecium and is actually a rare example in Serbian of a concept which (besides the scientific term which originates from the Latin *paramecium*) also has a more informal term, *papučica.* This term is widely used and it can be found in many elementary and high schools textbooks.

More often it is the case that in the Princeton wordnet a synset consists of several literals while its corresponding Serbian synset contains one or at most two literals. A good example is the Serbian adjective *amebni***,** for which as much as five synonyms exist in English {amoebic**:**1, amebic:1, ameban:1, amoeban:1, amoebous:1}. It can be easily seen that all these synonyms are actually variant forms.

The practical work on the development of the Serbian wordnet was, in many cases, a true research task for which I had to use traditional resources, such as printed dictionaries and textbooks. After consulting these, I looked for confirmation on Internet, either by using Google or by consulting reliable academic, professional or educational web sites. This kind of research task can be illustrated by the synset that refers to the well-studied hereditary disease hemophilia. There are different types of hemophilia depending on the type of the genetic disorder of clotting that cause it. One type is *hemophilia B***,** for which in English the synonymous term *Christmas disease* is also used. At first it seemed to me that the synonymous term is not known in Serbian or that it can be translated as *Božićna bolest*, i.e. the disease connected to *Christmas***.** I have, however, consulted the classical textbook "Genetics in Medicine" by Kičić (Kičić & Krajičanić, 1989)

in which he speaks about *Kristmasova bolest*, the disease of the boy *Christmas*. So, finally, the synsets in English and Serbian wordnets that correspond to this disease look like this:

<SYNSET> <ID>ENG20-13364794-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>hemophilia B <SENSE>1</SENSE> </LITERAL> <LITERAL>haemophilia B <SENSE>1</SENSE> </LITERAL> <LITERAL>Christmas disease <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-13364162-n <TYPE>hypernym </TYPE> </ILR> <DEF>a clotting disorder similar to hemophilia A but caused by a congenital deficiency of factor IX</DEF> <DOMAIN>genetics</DOMAIN> <SUMO>DiseaseOrSyndrome <TYPE>+</TYPE> </SUMO> </SYNSET>	<SYNSET> <ID>ENG20-13364794-n </ID> <SYNONYM>- <LITERAL>hemofilija B <SENSE>1</SENSE> </LITERAL> <LITERAL>Kristmasova bolest <SENSE>1</SENSE> </LITERAL> </SYNONYM> <DEF>poremećaj zgrušavanja veoma sličan hemofiliji A ali se javlja usled kongenitalne deficijencije faktora IX (po imenu dečaka Christmas, prvi put dijagnostikovana 1952. godine)</DEF> <SNOTE>Uradila S.Antonić, postdiplomac C. Krstev</SNOTE> <POS>n</POS> <ILR>ENG20-13364162-n <TYPE>hypernym</TYPE> </ILR> </SYNSET>

In the course of my practical work I used many general dictionaries: English-Serbian dictionary (Benson, M. 1997), "The New Merriam-Webster Dictionary" (Merriam-Webster 1989), The Dictionary of the Serbo-Croatian Literary Language (Matica srpska i Matica hrvatska, 1967) and The Dictionary of Foreign Words and Phrases (Klajn, I. & Šipka, M. 2006).

Since general dictionaries did not contain many of the technical terms from the scientific domain that I have processed, I had to use many textbooks as well. For microbiology, I used textbooks (Jemcev, V. T. & Đukić, D. 2000), (Tešić, Ž. & Todorović, M. 1992), (Jarak, M. & Govedarica, M. 2003), and for the genetics (Marinković, D. et al, 1989), (Kičić, M. & Krajičanić, B. 1989) and (Dumanović, J. et al, 1985). Besides these, I also used reference books on invertebrate zoology (Krunić, M. 1990), cytology (Grozdanović-Radovanović, J. 1985), histology (Grozdanović-Radovanović, J. 1980), and the development of animals (Ćurčić, B. 1985). Work on the synsets related to viruses was very interesting, but also demanding, as they belong to the class of the most simple microscopic organisms that mutate easily and, due to this, they are difficult to research. An excellent textbook on medical virusology from Prof. Ljubiša Krstić (Krstić, Lj. 2005) was very useful as a most reliable source for the terminology connected to viruses.

Useful solutions to the problems that occurred during the production of the synsets from the domain of biology for the Serbian wordnet were sometimes unexpectedly found on the Internet sites Wikipedia and Vokabular. Useful Internet resources were also the dictionary on the web site of the Faculty of Philosophy of Novi Sad. I also used a site about the protection of plants and a site on Human Genome Project during my work in 2006 (although it is not active any more).

4.3. The Domain of Religion

Nevena Ivković-Berček

Since I am a graduate student of theology and I work as a librarian in the Faculty of Theology's Library in Belgrade, I am familiar with the terms related to religion and religious literature, many of which I acquired during my studies and I encounter on a regular basis in the course of my present work.

My task in this project was to adopt synsets from the domain of religion, from the Princeton wordnet to the Serbian wordnet. The processed subsets of synsets are comprised of synsets that are connected to the SUMO ontological categories *Religious Process* (130 synsets) and *Religious Organizations* (160 synsets).

One typical example of the synsets from the Princeton and the Serbian wordnet that are connected to the ontological category *Religious Process* looks as follows:

<SYNSET> <ID>ENG20-00979294-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>anointing of the sick <SENSE>1</SENSE> </LITERAL> <LITERAL>extreme unction <SENSE>1</SENSE> </LITERAL> <LITERAL>last rites <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-00974693-n <TYPE>hypernym </TYPE> </ILR> <DEF>a Catholic sacrament; a priest anoints a dying person with oil and prays for salvation</DEF> <DOMAIN>religion</DOMAIN> <SUMO>ReligiousProcess <TYPE>+</TYPE> </SUMO> </SYNSET>	<SYNSET> <ID>ENG20-00979294-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>poslednja pričest <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR>ENG20-00974693-n <TYPE>hypernym </TYPE> </ILR> <DEF>katolička sveta tajna; sveštenik miropomazuje umiruću osobu uljem i moli se za spasenje</DEF> <SNOTE>Uradila N. Ivković, postdiplomac C. Krstev</SNOTE> </SYNSET>

I encountered two types of problems during my preparation of synsets related to religious activities and organizations. It is well known that in Serbian many terms known in foreign languages simply do not exist. I have encountered this problem in trying to find names for the religious avtivities unknown to our culture or for religious communities that do not exit in our society. For instance, the Princeton wordnet contains the synset {popery:1} (an offensive term for the practices and rituals of the Roman Catholic Church) which was connected to the ontological category *Religious Process*. This concept is not recognized in our language and it is not used.

Among religious organizations represented in the Princeton wordnet is the Protestant denomination founded by Mary Baker Eddy in 1866, which is either called *Christian Science* or *The Church of Christ Scientist*. This denomination does not exist in our society and therefore the name for it in Serbian does not exist either. After consultations with Prof. Krstev, I decided that the most appropriate term in Serbian would be "Crkva Hristovih naučnika".

I have used in my work many lexicons, dictionaries and encyclopedias. First of all I have used the studies: *The Attack on Religious Freedom* (Bjelajac, B. & Vidović, D. 2001), *Religious Sects and Politics* (Branković, T. 2000), *Religious Sects* (Đurđević-Stojković, B. 2002), *The Encyclopedia of Living Religions* (Krim, K. 1992) and *A Dictionary of Biblical Theology* (Leon-Dufour, X. 1969). When the scientific and technical literature did not offer a translation in Serbian or explanation for some terms, I had to use general lexicons of foreign words (Anić, Š. & Klaić, N. 2002; Vujaklija, M. 2005; Klaić, B.1951; Klajn, I. & Šipka, M. 2006) an English language dictionary (MacMillan, 2002), a Serbian language dictionary (Moskovljević, M. 1966) and a bilingual English-Serbian dictionary (Ristić, S. et al, 1956).

The Internet sources that I have often used and that helped me a great deal in my work were the Merriam Webster web site, both the Serbian and English versions of Wikipedia, Vokabular, (which was conceived as a free service that offers a general Serbian-English dictionary and other services), and *Metak* (English: bullet) as a separate dictionary service of the web site SerbianCafe. Besides these, I also used different search engines, but most of the useful information I found by Google and Krstarica, the Serbian search engine.

4.4. The Domain of Literature

Zorica Zorica

I graduated from the Department of Serbian Language and Literature at the Faculty of Philosophy in Novi Sad, and now I work as a librarian in the Legacy of the collector Rajko Mamuzić Art Gallery in Novi Sad.

My task in this project was to adopt synsets from the domain of literature from the Princeton wordnet for the Serbian wordnet. This included

the synsets that were labeled in the Princeton wordnet as belonging to the domain "literature". Most of the processed synsets were connected to the ontological categories *Text* (218), *Writing* (17), *Linguistic Expression* (16), *Content Development* (11) while the others belonged to some other categories related to the theory of literature and rhetoric. The total number of processed synsets was 355. Here is one typical example of an adopted synset, which is connected to the ontological category *Text*:

<SYNSET>	<SYNSET>
<ID>ENG20-05976529-n	<ID>ENG20-05976529-n
</ID>	</ID>
<POS>n</POS>	<POS>n</POS>
<SYNONYM>	<SYNONYM>
<LITERAL>fable	<LITERAL>basna
<SENSE>2</SENSE>	<SENSE>2</SENSE>
</LITERAL>	</LITERAL>
<LITERAL>parable	<LITERAL>parabola
<SENSE>1</SENSE>	<SENSE>1</SENSE>
</LITERAL>	</LITERAL>
<LITERAL>allegory	<LITERAL>alegorija
<SENSE>1</SENSE>	<SENSE>1<SENSE>
</LITERAL>	</LITERAL>
<LITERAL>apologue	<LITERAL>apolog
<SENSE>1</SENSE>	<SENSE>1<SENSE>
</LITERAL>	</LITERAL>
	<LITERAL>poučna priča
</SYNONYM>	
<ILR>	<SENSE>1<SENSE>
<TYPE>hypernym	</LITERAL>
</TYPE>	<LITERAL>poučna basna
ENG20-05974336-n	
</ILR>	<SENSE>1<SENSE></LITERAL>
	</SYNONYM>
<ILR>	<ILR>
	<TYPE>hypernym
	</TYPE>
	ENG20-05974336-n
	</ILR>
<DEF>a short moral story (often with animal characters)</DEF>	DEF>kratka poučna priča u kojoj su često junaci životinje</DEF>
<DOMAIN>literature	
</DOMAIN>	<SNOTE>Uradila Z.
<SUMO>Text	Zorica, postdiplomac C.
<TYPE>+</TYPE>	Krstev</SNOTE>
</SUMO>	
</SYNSET>	</SYNSET>

During my work on the development of the Serbian wordnet I encountered several kinds of problems. One of the main problems was ascertaining what the appropriate term, and which was not a neologism, in Serbian was for some of the concepts represented in the Princeton wordnet.

I also faced another type of problem when I tried to find appropriate terms for concepts that are not lexicalized in Serbian or for which some more general terms are used. Some synsets, such as {sloganeer:1} (coin new slogans) or {novelization:1, novelisation:1} (converting something into the form of a novel), I left untranslated. The synset {potboiler:1} (a literary composition of poor quality that was written quickly to make money (to boil the pot)) also was left untranslated, since a precise term for this notion does not exist in Serbian literary terminology. The closest notions are *trivijalna književnost* or *petparački roman*. These, however, do not represent the same concept since "potboiler" does not have to be a novel, while trivial literature need not necessarily be written for the purpose of quick profits.

The third type of problem emerged from terms like {novelette:1, novella:1}. In Anglo-Saxon it has in literature terminology the meaning of a short novel, while similar terms in Serbian literature terminology have the meaning "a very short novel" and "a kind of a story". I have translated these terms using the Anglo-Saxon literature terminology and I have put an additional note explaining this in the gloss. I have applied the same approach to the synset {romance:5} (a novel dealing with idealized events remote from everyday life) since the term *romansa* has the similar meaning in our literature terminology. However, one has to keep in mind that the more usual meaning for this term in Serbian is "epic or lyric folk song". Similarly, the direct translation in Serbian for the literal "story" from the synset {story:2} (a piece of fiction that narrates a chain of related events) would be *priča*; in the Serbian literature terminology, although, the term *pripovetka* is more often used to denote that type of a prose work.

I have used in my work many lexicons and dictionaries: A Lexicon of Literary Terms (Živković, D. 1992), Lexicons of Foreign Words and Phrases (Vujaklija, M. 1986) as well as (Klajn, I. &

Šipka, M. 2007.), The Terminological Dictionary of Librarianship (Kovačević, Lj. 2004), and The Textbook on the Theory of Literature and the Theory of Literacy (Živković, D. 2001).

Internet resources that I found to be the most useful were Vokabular.org, Rastko.org and the search engines Google, Yahoo and Krstarica. All of these, however, proved that electronic resources related to Serbian literature terminology are very scarce and insufficient.

4.5. The Domain of Law

Vesna Crnogorac

As a graduate of law who has worked five years as a lawyer, I am familiar with the basic legal terms. As well, for the last ten years I have worked as a librarian performing different tasks, and for the past four years I have also worked as a journalist. Since 2006 I have been a professional secretary for the Association of the Librarians of Serbia.

My task in this project was to adapt synsets from the domain of law from the Princeton wordnet for contribution to the Serbian wordnet. This included synsets that were in the Princeton wordnet labeled as belonging to the domain "law". The selected judicial terms were related to most of the law branches that are in use in our judicial system: criminal and criminal procedural law, obligational law, commercial and international commercial law, civil and civil procedural law, administrative law, public international law, inheritance law, and penology. Among them, however, were also some basic terms that are usually acquired in the scope of the course Introduction to Law or something similar. The Princeton wordnet contains many synsets that belong to the domain 'law', and since it was not possible to process them all on this occasion, the following criteria was applied for selection: (a) a selected synset belongs to the so-called third set of basic concepts established in the Balkanet project and has not yet been included in the the Serbian wordnet (42 synsets); or (b) it is connected to the SUMO ontological category 'Certificate' (73 synsets). The total number of synsets processed in this way was 115.

A typical example of the synsets from the Princeton and the Serbian wordnet that belongs to criminal procedural law looks like this:

```
<SYNSET>
  <ID>ENG20-01122850-n
 </ID>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>conviction
      <SENSE>2</SENSE>
    </LITERAL>
    <LITERAL>judgment of
conviction
      <SENSE>1</SENSE>
   </LITERAL>

<LITERAL>condemnation
      <SENSE>5</SENSE>
    </LITERAL>
    <LITERAL>sentence
      <SENSE>2</SENSE>
    </LITERAL>
    </SYNONYM>


<ILR><TYPE>hypernym
  </TYPE>
       ENG20-01122569-n
 </ILR>
    <ILR><TYPE>near_an-
tonym</TYPE>
       ENG20-01127432-n
 </ILR>
    <ILR><TYPE>eng_deri-
vative</TYPE>
        ENG20-00876567-v
 </ILR>
    <ILR><TYPE>eng_deri-
vative</TYPE>
        ENG20-00876935-v
 </ILR>
    <ILR><TYPE>category_
domain</TYPE>
        ENG20-06135956-n
 </ILR>
    <DEF>(criminal law) a
final judgment of guilty in a
criminal case and the punish-
ment that is imposed</DEF>

    <USAGE>the conviction
came as no surprise</USA-
GE>
     <BCS>3</BCS>
    <DOMAIN>law</DO-
MAIN>
    <SUMO>Sentencing
       <TYPE>=</TYPE>
   </SUMO>
</SYNSET>
```

```
<SYNSET>
    <ID>ENG20-01122850-n
 </ID>
     <POS>n</POS>
     <SYNONYM>
     <LITERAL>osuđujuća
presuda
       <SENSE>5</SENSE>
    </LITERAL>


     </SYNONYM>

<ILR><TYPE>hypernym
  </TYPE>
       ENG20-01122569-n
 </ILR>
     <ILR><TYPE>near_
antonym</TYPE>
        ENG20-01127432-n
 </ILR>
     <ILR><TYPE>eng_
derivative</TYPE>
        ENG20-00876567-v
 </ILR>


<ILR><TYPE>category_
domain</TYPE>
        ENG20-06135956-n
 </ILR>

     <DEF>(krivično pravo)
pravnosnažna,osuđujuća
presuda u krivičnom
postupku sa dosuđenom
kaznom
    </DEF>
     <USAGE>Za primanje
mita optuženo je 46 lica , a za
davanje mita 35. Osuđujuća
presuda za primanje mita
doneta je u 31 slučaju , dok je
za davanje mita osuđeno 30
osoba</USAGE>
     <BCS>3</BCS>
     <DOMAIN>law</DO-
MAIN>
     <SUMO>Sentencing
        <TYPE>=</TYPE>
   </SUMO>
</SYNSET>
```

The work on the adaptation of synsets for the law domain of the Serbian wordnet was very interesting and it initiated my professional curiosity to learn more. Despite this fact I faced problems all the time that derived from the fact that the Serbian legal system and the American legal system are very different. Namely, the American legal system belongs to the so-called Anglo-American law, while the Serbian system belongs to Euro-continental law. Therefore, the Serbian legal system either does not use or defines many law terms existing in the Princeton wordnet differently. For this reason I had to consult several sources in order to find the most adequate solution for the concepts that are not well-known in the Serbian legal system. One example that illustrates this is given below:

```
<SYNSET>
  <ID>ENG20-06150174-n
 </ID>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>assize
      <SENSE>2</SENSE>
  </LITERAL>
  </SYNONYM>
  <ILR>ENG20-06149686-n
 <TYPE>hypernym
</TYPE>
  </ILR>
  <ILR>ENG20-07928837-n
    <TYPE>category_do-
main</TYPE></ILR>

  <DEF>an ancient writ
issued by a court of assize to
the sheriff for the recovery of
property</DEF>

   <DOMAIN>law
  </DOMAIN>
  <SUMO>Certificate
  <TYPE>+</TYPE>
 </SUMO>

</SYNSET>
```

```
<SYNSET>
   <ID>ENG20-06150174-n
 </ID>
   <POS>n</POS>
   <SYNONYM>
      <LITERAL>uredba
       <SENSE>2</SENSE>
   </LITERAL>
   </SYNONYM>
   <ILR>ENG20-06149686-n
<TYPE>hypernym</TYPE>

   </ILR>
   <ILR>ENG20-07928837-n

<TYPE>category_domain
</TYPE></ILR>
   <DEF>sudski nalog
(izdavan u prošlosti)
administrativnom činovniku
za povraćaj imovine </DEF>
   <NOTE>Not-lexicalized
in Serbian</NOTE>
   <DOMAIN>law
  </DOMAIN>
   <SUMO>Certificate
   <TYPE>+</TYPE>
  </SUMO>
</SYNSET>
```

The XML element of the synset description <NOTE> contains a note that indicates that this concept is not actually lexicalized in Serbian. It is interesting to note that in the Bulgarian wordnet, which was also being developed during the Balkanet project, this concept is not lexicalized either. Therein the content of the element <LITERAL> in the Bulgarian wordnet is ‘разпореждане, издавано на шерифа от съда за възстановяване на собственост:1’.

During my practical work I regularly consulted the dictionary of legal terms (Jovanović, J. & Todorović, S. 2004). Since general lexicons do not generally contain professional terms form the law domain, I also used many textbooks form this domain (Kovačević Kuštrimović, R. 1997, Jovanović, Lj. 2000, Stanković, G. 1998).

Since Prof. Krstev and I had the impression that the electronic texts from the law domain are better represented then e-texts from other domains, we decided to verify the suggested solutions by consulting the corpus of contemporary Serbian, and when it did not offer the confirmation needed, through searching the Internet. The examples found were included in the <USAGE> element of the XML description of synsets, as can be seen in the given examples of this subsection. The search for confirmation proved to be a very delicate task, as we wanted to find the exact term and not its paraphrase. The synset {potvrđena presuda:1} that corresponds to the synset in the Princeton wordnet {affirmation:4} (a judgment by a higher court that the judgment of a lower court was correct and should stand) illustrates this point. An example of usage found in the Serbian corpus was “Kaznu od 12 godina zatvora Nenadu B. Vrhovni sud je preinačio na 15 godina robije, dok je Milu I. <potvrđena presuda> od tri i po godine zatvora” (The Supreme Court has reversed the sentence of Nenad B. to 15 years in prison, from 12 years of penal servitude, while for Milo I. the sentence was affirmed for three and a half years in prison.). Maybe it cannot be said that this example does not confirms the existance of the term *potvrđena presuda*, but one has to note that in the given example it is not a term with the syntactic structure *adjective+noun* that is used, but rather a verb with its complement. Another interesting example is the synset {speeding ticket:1} (a ticket issued for driving above the speed limit). The first offered solution was *kazna za prekoračenje brzine*, but a search on the Internet showed that its variant form *ka-*

zna zbog prekoračenja brzine can also be used, as in the example "Dvojica saobraćajnih policajaca naplatila su ministru unutrašnjih poslova Srbije Boži P. kaznu zbog <prekoračenja brzine> na Ibarskoj magistrali" (Two traffic officers issued a fine to the Serbian minister of the internal affairs Boža P. for driving above the speed limit on the Ibar highway). Thereafter this variant term was added to the Serbian synset.

4.6. The Domain of Librarianship and Publishing

Ljiljana Macura

I graduated from the Library and Information Sciences Department at the Faculty of Philology in Belgrade. After many years working in different types of libraries, I now work as a librarian and informatician in the Serbian National Library's Department for Information Services, in the scientific and information center. During these years, from the beginning of my studies of library and information sciences to this very moment, my perspective through which I have considered these terms has been continuously expanding.

I processed 62 synsets from the domain 'publishing' that are connected to the ontological category 'Book' and 20 more synsets that belong to various domains but are related to catalogues in general.

The notion *catalogue* especially draws the attention of librarians. In English *catalogue* can be a noun, a verb, and can be used in the formation of an adjective. In Serbian, it exists as many types of words: *katalog*, *katalogizacija* (nouns), *katalogizirati* (verb), *kataloški*, *katalogiziran* (adjectives). The forms *katalogizirati* and *katalogizirano* are more specific to the western variant of Serbo-Croatian, but they are also used regularly in Serbian as well. These are used in the Serbian language, both in general and in a narrowed sense: *kataloški obraditi/obrađeno* (process/processed for a catalogue), *uneti/uneto u katalog* (enter/entered into catalogue), *sačiniti katalog* (to make a catalogue), etc.

In the Princeton wordnet we located numerous synsets consisting of several literals to which corresponded synsets with only one literal in Serbian and vice versa. The following examples illustrate this:

- The synset with only one literal {kuvar:1} corresponds to the synset {cookbook:1, cookery book:1} (a book of recipes and cooking directions) from the Princeton wordnet;
- The synset with several literals {knjiga narudžbi:1, knjiga porudžbi:1, knjiga trebovanja:1} corresponds to the synset {order book:1} (a book in which customers' orders are entered; usually makes multiple copies of the order) from the Princeton wordnet;
- The synset with several literals {broširano izdanje:1, knjiga u mekom povezu:1} corresponds to the synset {paperback book:1, paperback book:1, paperback:1, softback book:1, softback:1, soft-cover book:1, soft-cover:1} (a book with paper covers) from the Princeton wordnet.

The example of two corresponding synsets:

< SYNSET > <ID>ENG20-06015176-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>tome <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR> <TYPE>hypernym </TYPE> ENG20-06013091-n </ILR> <DEF>a (usually) large and scholarly book knjiga (obično) velika, i za obrazovanog korisnika </DEF> <DOMAIN>publishing </DOMAIN> <SUMO>Book <TYPE> +</TYPE></SUMO> </SYNSET>	< SYNSET > <ID>ENG20-06015176-n </ID> <POS>n</POS> <SYNONYM> <LITERAL>tom <SENSE>1</SENSE> </LITERAL> <LITERAL>sveska <SENSE>1</SENSE> </LITERAL> <LITERAL>svezak <SENSE>1</SENSE> </LITERAL> <LITERAL>knjiga <SENSE>1</SENSE> </LITERAL> </SYNONYM> <ILR> <TYPE>hypernym </TYPE> ENG20-06013091-n </ILR> <DEF>knjiga (obično velika), i za obrazovanog korisnika </DEF> < SNOTE>Uradila Ljiljana Macura, postdiplomac C. Krstev< /SNOTE> </SYNSET>

Some examples of concepts not lexicalized in Serbian are:

- There is no equivalent translation in Serbian for the synset {grimoire:1} (a manual of black magic (for invoking spirits and demons)). Therefore, a descriptive literal is used in the Serbian wordnet {knjiga magije:1};
- Neither the synset {consuetudinary:1, consuetudinal:1} (a manual describing the customs of a particular group (especially the ceremonial practices of a monastic order)) has an adequate translation in Serbian, although the term *običajnik* would be very good if this concept were known in Serbian**;**
- The phrase *zbirka basni* is a relatively adequate translation for the synset {bestiary:1} (a medieval book (usually illustrated) with allegorical and amusing descriptions of real and fabled animals) since there is no literal translation in Serbian.

For practical reasons, but also out of habit, most of the time I consulted dictionaries and lexicons in paper form (Institut za strane jezike, 2005), (Benson, M. 1997), (Simić, D. 2005) (Vukičević, B. 2001), (JLZ, 1974), (Vujaklija, M. 2005), (Matica srpska i Matica hrvatska, 1967). I also used a terminological dictionary for librarians in electronic form (BTR ONLINE), and the Princeton wordnet itself. Of course, I have also used the Internet as a general source of information, but I think that traditional resources are still more reliable.

5. Conclusion

The Serbian wordnet today has 14,593 synsets of which 2,240 were prepared by the participants in this cooperative work. We think that our common efforts have shown that cooperative work in the development of the Serbian wordnet is possible. Our present experience, although, has also indicated it is necessary that one person act as an editor in order to coordinate the work of all collaborators and to control all prepared synsets before their definite inclusion into the database. It should be kept in mind that most of the collaborators have invested their specific knowledge, valuable experience in information retrieval and usage of reference literature, as well as great enthusiasm, into this common project; however, many of them, quite naturally, are not familiar with linguistics, especially computational linguistics, and because of this, the role of an editor is crucial. We expect that a new web tool for development and browsing of wordnets, particularly the Serbian wordnet, that is being developed on the bases of ideas presented in (Obradović, I. & Stanković, R. 2008) will help facilitate the work of both collaborators and editor in the future.

The automatic construction of wordnets presents a topic that has been relevant for many years and has attracted much research. Several papers presented at the *Global Wordnet* Conference, which was held in Szeged in January 2008, presented results about the automatic construction of wordnets for particular languages (Slovenian, Polish, etc). The general evaluation of automatically produced wordnets is given in the paper (De Mello, G. & Weikum, G. 2008). The authors of this paper concluded that the automatically produced wordnets are not only useful in many computational applications, but also for the production of traditional lexical resources. It should be kept in mind that the automatic construction of a wordnet usually relies heavily on the usage of multilingual language resources, especially textual resources, in digital form. The experience we have had in the development of the Serbian wordnet shows that the availability of textual resources in digital form for some domains is insufficient even for human work, and that they would probably be quite unusable for the automatic generation of synsets from these domains. Moreover, even the coverage of some domains with traditional lexicons and manuals is in many cases completely inadequate.

[1]The terms "prvi jezik" and "maternji jezik" were found in the Serbian translation of the Cambridge Encyclopedia of Languages (Kristal, 1995).

Used reference material: Dictionaries, Textbooks and Manuals

Anić, Šime. Klaić, Nikola. Domović, Želimir. *Rječnik stranih riječi i izraza*. Zagreb: Sani-Plus, 2002.

Bjelajac, Branko, Vidović, Dane. *Udar na verske slobode*. Beograd: Alfa i Omega, 2001.

Benson, Morton. *Englesko-srpskohrvatski rečnik*. Prosveta : Ljubljana 1980

Benson, Morton. *Englesko-srpski rečnik* [Elektronski izvor]. (1997)

Branković, Tomislav. *Sekte i politika*. Despotovac: Narodna biblioteka "Resavska Škola", 2000.

Bugarski, Ranko. *Uvod u opštu lingvistiku*. Beograd: Čigoja, 1996

Bugarski, Ranko. *Jezici*. Beograd: Čigoja, 1996

Ćosić, Pavle. *Rečnik sinonima i tezaurus srpskog jezika*. Beograd: Kornet, 2007

Ćurčić, Božidar. *Razviće životinja*, Naučna knjiga, Beograd. (1985)

Dumanović, Janko, Marinković, Dragoslav, Denić, Miloje. *Genetički rečnik*, Naučna knjiga, Beograd. (1985)

Đorđević, Radmila. *Gramatika engleskog jezika.* Beograd: Izdanje autora, 2002

Đurđević-Stojković, Biljana. *Verske sekte*. Beograd: Narodna knjiga, 2002.

Englesko-srpski srpsko-engleski rečnik sa gramatikom =English-serbian serbian-english Dictionary & Grammar : ESSE. Beograd: Institut za strane jezike, 2005

Grozdanović-Radovanović, Jelena. *Citologija*, Naučna knjiga, Beograd. (1985)

Grozdanović-Radovanović, Jelena. *Histologija*, Beograd. (1980)

Jarak, Mirjana, Govedarica, Mitar. *Mikrobiologija*, Poljoprivredni fakultet, Novi Sad. (2003)

Jemcev, Vsevolod Tihonovič, Đukić, Dragutin. *Mikrobiologija*, Vojnoizdavački zavod, Užice. (2000)

Jovanović, Ljubiša. *Krivično pravo*. Niš: Pravni fakultet. (2000).

Jovanović, Jasmina i Todorović, Svetlana. *Rečnik pravnih termina:srpsko – englesko-francuski =Legal Dictionary: English – Serbian = Termes juridigues : francais – serbe*, Savremena administracija, Beograd. (2004).

Kičić, Miroljub, Krajičanić, Branka. *Medicinska genetika*, Zavod za udžbenike i nastavna sredstva, Beograd. (1989)

Klaić, Bratoljub. *Rječnik stranih riječi, izraza i kratica*. Zagreb: Državno izdavačko poduzeće Hrvatske, 1951.

Klajn, Ivan, Šipka, Milan. *Veliki rečnik stranih reči i izraza*. Novi Sad: Prometej, 2006.

Klajn, Ivan, Šipka, Milan. *Veliki rečnik stranih reči i izraza*. Novi Sad: Prometej, 2007

Kovačević, Ljiljana. *Bibliotekarski terminološki rečnik: englesko-srpski, srpsko-engleski.* Beograd: Narodna biblioteka Srbije, 2004.

Kovačević Kuštrimović, Radmila. *Građansko pravo*. Niš: Pravni fakultet, 1997.

Krim, Kit. *Enciklopedija živih religija*. Beograd: Nolit, 1992.

Kristal, Dejvid. *Kembrička encikopedija jezika.* Beograd: Nolit, 1995

Kristal, Dejvid. *Enciklopedijski rečnik moderne lingvistike*. Beograd: Nolit, 1998

Krstić, Ljubiša. *Medicinska virusologija*, Grafopan, Beograd. (2005)

Krunić, Miloje. *Zoologija invertebrata. Deo 1*, Naučna knjiga, Beograd. (1990)

Leksikon JLZ. Jugoslavenski leksikografski zavod, Zagreb 1974

Leon-Dufour, Xavier. *Rječnik biblijske teologije*. Zagreb: *Kršćanska sadašnjost*, 1969.

MacMillan English Dictionary for Advanced Learners: International Student Edition. Oxford: Bloomsberry Publishing Plc, 2002.

The New Merriam-Webster Dictionary. Merriam-Webster, Springfield. (1989)

Marinković, Dragoslav, Tucić, Nikola, Kekić Vladimir. *Genetika*, Naučna knjiga, Beograd. (1989)

Mihailović, Ljiljana. *Upotreba pasivnih glagolskih oblika u savremenom engleskom jeziku*. Filološki fakultet Beogradskog univerziteta,Monografije, knjiga XII, Beograd 1967

Moskovljević, Miloš. *Rečnik savremenog srpskohrvatskog književnog jezika*. Beograd: Tehnička knjiga-Nolit, 1966.

Putanec, Valentin *Francusko-hrvatski ili srpski rječnik.* Zagreb: Školska knjiga, 1974.

Rečnik srpskohrvatskoga književnoga jezika. Novi Sad : Zagreb, Matica srpska i Matica hrvatska, 1967-1976

Ristić, Svetomir. Simić, Živojin. Popović, Vladeta. *Enciklopediski englesko-srpskohrvatski rečnik*. Beograd: Prosveta, 1956.

Sapir, Edvard. *Jezik*. Novi Sad: Dnevnik, 1992

Simić, Dušan. *Englesko-srpski enciklopedijski rečnik*. Kragujevac: Centar za naučna istraživanja SANU i Univerziteta : DSP, 2005

Stanković, Gordana.*Građansko procesno pravo*. Niš: Pravni fakultet. (1998)
Tešić, Živojin, Todorović Milan. *Mikrobiologija*, Naučna knjiga, Beograd. (1992)
Vujaklija, Milan. *Leksikon stranih reči i izraza*. Beograd: Prosveta, 1980
Vujaklija, Milan. *Leksikon stranih reči i izraza*. Beograd: Prosveta, 1986
Vujaklija, Milan. *Leksikon stranih reči i izraza*. Beograd: Prosveta, 2005.
Vukičević, Branko. *Pravni rečnik : englesko-srpski sa obrascima pravnih*
akata : 40.000 terminoloških jedinica. Beograd : Grmeč-Privredni pregled 2001.
Živković, D. (urednik): *Rečnik književnih termina.* Beograd: Nolit, 1992.
Živković, D. *Teorija književnosti sa teorijom pismenosti – priručnik za nastavnike i učenike*, Beograd: Draganić, 2001.

Internet Sources

BTR ONLINE (a web edition of the terminological dictionary for librarians : English/Serbian and Serbian/English) (http://btr.nbs.bg.ac.yu/)
Danko Šipka – a mailing list for Serbian as a foreign language (http://www.public.asu.edu/~dsipka/F2.TXT)
Google (www.google.com)
Human Genome Project Information (http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml)
The corpus of contepmorary Serbian (http://www.korpus.matf.bg.ac.yu)
Krstarica (www.krstarica.com)
Metak (http://www.metak.com)
Miriam Vebster (www.merriam-webster.com/dictionary)
PhraseBASE (http://www.phrasebase.com/croatian/languages/index.php?cat=18)
PROZ - T (http://srp.proz.com/job?&print=1)
RASTKO Project – An Internet library of Serbian culture (http:www.rastko.org.yu)
The Dictionary of the Faculty of Philosophy in Novi Sad (used during 2006) (http://www.ff.ns.ac.yu/elpub/specst/recnik.htm#recnik)
ST-L (a mailing list for Serbian terminology) (http://www.staff.amu.edu.pl/~sipkadan/korpus.html)
Wikipedia (http://sr.wikipedia.org, http://en.wikipedia.org, http://hr.wikipedia.org)
Vokabular (http://www.vokabular.org)
The protection of plants http://www.poljoprivreda.info/?oid=9&id=639
http://www.phrasebase.com/croatian/languages/index.php?cat=18

Literature:

Bentivogli, L., P. Forner, B. Magnini and E. Pianta. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing, in COLING 2004 Workshop on "Multilingual Linguistic Resources", Geneva, Switzerland, August 28, 2004, pp. 101-108.
Christodulakis, D. N. (ed.). (2004). Design and Development of a Multilingual Balkan Wordnet (BalkaNet IST-2000-29388) – Final Report.
Fellbaum, C., ed. (1998). WordNet: An Electronic Lexical Database. Cambridge: Mass: MIT Press.
Krstev C. (2006). Specifični koncepti Balkana u semantičkoj mreži Wordnet. U Zborniku radova "Susreti kultura", Novi Sad, decembar 2004, eds. Ljiljana Subotić et al, pp. 275-285, Novi Sad: Univerzitet u Novom Sadu, Filozofski fakultet.
Magnini, B. and G. Cavaglià. (2000). Integrating Subject Field Codes into WordNet. In Gavrilidou M. et al. (Eds.) Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens: Greece, 31 May – 2 June, 2000, pp. 1413-1418.
De Mello, G. and Weikum.G. (2008). On the Utility of the Automatically Generated Wordnets. In Proceedings of the Fourth Global WordNet Conference. Syeged: Hungary, January 22-25, 2008, pages 147-161.
Miller, George A. (1990). Nouns in Wordnet: A Lexical Inheritence System. Journal of Lexicography 3(4): pp. 245-264.
Niles, I. and Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE '03).Las Vegas: Nevada, June 23-26.
Obradović, I. , Stanković, R. (2008). Software tools for Serbian lexical resources. Infotheca (this number)
Pease, A., and Niles, I. (2002). IEEE Standard Upper Ontology: A Progress Report. Knowledge Engineering Review, Special Issue on Ontologies and Agents. 17, 65-70.
Stamou, S., Nenadić, G., Christodulakis, D. (2004). Exploring BalkaNet Shared Ontology towards Multilingual Conceptual Indexing. In Proceedings of the 4th Language Resources and Evaluation Conference (LREC). Lisbon: Portugal.
Tufiş, D., Cristea, D., Stamou, S. (2004). Balkanet: Aims, Methods, Results and Perspectives. A General Overview in Special Issue on BalkaNet Project, Roma-

nian Journal on Information Science and Technology, Tufiş, D. (ed.). Bucureşti: Publishing house of the Romanian academy, pages 9-43.
Tufiş, D. and Svetla K. (2007). Ontology-Supported Text Classification Based on Cross-Lingual Word Sense Disambiguation. In WILF 2007, LNAI 4578, Masulli, F., Mitra, S. and Pasi, G. (eds.). Berlin Heidelberg: Soringer-Verlag, pages 447-455.
Vossen, P. (2004) Eurowordnet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual-Index, International Journal of Lexicography, 17(2), pages 161-173.
Vossen, P. (2004) Introduction to the Special Issue on the BalkaNet Project. In Special Issue on BalkaNet Project, Romanian Journal on Information Science and Technology, Tufiş, D. (ed.). Bucureşti: Publishing house of the Romanian academy.