SOUTHERN JOURNAL OF LINGUISTICS
29: 106-118 • © 2007 Southeastern Conference on Linguistics

An Approach to the Development of Language Specific Concepts in Wordnets*

Cvetana Krstev
Ivan Obradović
Duško Vitas
University of Belgrade

This article outlines the methodology developed for the selection and acquisition of concepts specific for the Balkans within the BalkaNet project. The main goal of this project was the construction of a multilingual semantic lexical database of the WordNet type for Balkan languages. Balkan specific concepts are concepts which are commonly used in the Balkan area, but are not recognized by the Princeton WordNet, the paradigm for the construction of wordnets for many languages. Domain and language characteristics of Balkan specific concepts are being discussed, as well as the establishment of the potential equivalents and the encountered problems in the application of the methodology adopted. The contribution of Serbian specific concepts to this subset is discussed in more detail.

1. Introduction

In 1985, George Miller, a renowned professor of psychology at Princeton University, and his associates from the Cognitive Science Laboratory started to develop the Princeton WordNet (PWN), or simply WordNet, a linguistic database that maps the way the mind stores and uses language. Its aim was to serve as some sort of a mental lexicon that can be used in the scope of psycholinguistic research projects (Fellbaum 1998). PWN was based on a semantic network of concepts, abstract ideas or mental symbols that denote all of the objects in a given category or class of entities, interactions, phenomena, or relationships between them. In PWN each concept is represented by a set of synonymous English word-sense pairs which, accompanied by a definition of

* This article is based on a presentation at the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages in Sofia, Bulgaria, in October, 2006.

the concept, form the synset for this concept. Concepts are interconnected by semantic relations, such as hypernym/hyponym ('kind of,' e.g. *animal/dog*) or holonym/meronym ('part of,' e.g. *hand/finger*). As of 2006, this database contains about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs.

The EuroWordNet project introduced multilingualism into the semantic network of concepts by building wordnets for seven European languages in a manner similar to PWN, and aligning them by interconnecting synsets representing the same concept in different languages by an Inter-Lingual-Index, or ILI (Vossen 1998). Along the same lines, the BalkaNet project 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 EuroWordNet (Tufiş 2004). Thirteen scientific and research institutions from Bulgaria, Greece, Romania, Serbia, Turkey, France, the Netherlands and Czech Republic gathered within the project consortium. Six teams were formed, each responsible for the development of a wordnet in one of the six languages.

The initial development of wordnets for the six BalkaNet languages was planned and realized synchronously. Namely, the core of each monolingual wordnet was built of several commonly agreed sets with a total of 8,516 concepts selected from PWN. Beyond these sets the network for each language has been developed independently, but always within the framework set by PWN. This approach generated some specific problems. Namely, during the work on the development of the network the following questions have 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 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.

2. Defining Balkan Specific Concepts

In an attempt to overcome some of the problems encountered, partners on the BalkaNet project agreed that one of the results of this project should be the incorporation of a set of concepts specific for Balkan languages in their wordnets.

Before the development of this set started, it had to be agreed what is to be a considered a concept specific for the Balkans, since various possible approaches were proposed. The decision had to be made, whether a Balkan specific concept (BSC) should be:

• a concept specific for a particular Balkan language (such as *кајмак* 'a milky spread made of skim' or *стара штедња* 'foreign currency saving accounts frozen by factual bankruptcy' for Serbian),

• a concept originating from one Balkan language which has spread to other Balkan and even European languages (such as *Атентат у Сарајеву* 'the assassination in Sarajevo'),

• a concept which is not necessarily specific for the Balkans only, but which is recognized as common in this area, while at the same time it has not been registered in PWN (for example, *пирамидална банка* 'banks offering extremely high interest rates' or *транзиција*[1] 'transition').

The first definition of a BSC has been rejected at the consortium level based on the conclusion that such a narrow determination would not be very productive. Although there were supporters of the idea that the set of BSCs should contain concepts specific for the Balkans only, the opinion prevailed that in view of future applications it would be more useful if the BalkaNet database contained the greatest possible number of concepts which are recognized as important in the Balkan area, regardless of their origin and dispersion.

The initial set of BSCs was formed according to the following procedure:

(1) Each team prepared a list of concepts specific for its language, cautiously verifying that the chosen concepts did not exist in PWN. Thus, for example, *баклава* 'baklava,' a natural candidate for a BSC, was not included, since it already existed in PWN.
(2) Each team compared its list prepared in the previous step with lists of concepts offered for other languages. The aim was to link similar specific concepts recognized in step one in different BalkaNet languages, thus forming a multilingual core of BSCs.
(3) After the core of BSCs was established, each team inspected the remaining concepts offered by other teams. If a concept, not proposed by a particular team, was nevertheless recognizable in the language the team was responsible for, the concept was added to the appropriate wordnet.

Once the initial set of BCSs was formed, each team was free to add other concepts, important for their language, to its wordnet. Needless to say, the rule determined in step one had to be observed, and other teams could always follow step three for newly emerging concepts.

In the first step the Serbian team offered 316 concepts (259 nouns, 9 verbs and 48 adjectives) which did not exist in PWN. The majority of them were related to food (*ајвар* 'pepper salad'), family relations (*јетрва* 'wife of one's husband's brother'), society—mainly its socialist heritage and the transition period (*ударник* 'a distinguished worker'), household (*кувариџа* 'embroidered cloth'[2]), religion (*Свети Сава* 'St. Sava'), customs (*слава* 'the day of the guardian saint'), mythology (*баук* 'an imaginary evil creature') and history (*Косовска битка* 'Kosovo battle'). Among adjectives, possessive adjectives derived from nouns which belong to the set of Serbian specific concepts dominated, such as. *ћевабџијин* 'belonging to *ћевабџија'* from *ћевабџија* 'one who produces and sells *ћевапчићи*[3]'. The verbs followed a similar pattern, e.g. *партизановати* 'act as a *партизан*' from *партизан* 'partisan.' All concepts defined in step one for Serbian were subsequently included in the Serbian WordNet (SWN).

Simultaneously, the other five teams independently defined concepts they considered to be specific for their languages. Thus 336 concepts were specified for Bulgarian, 309 for Greek, 545 for

Romanian, 332 for Turkish and 226 for Czech. Not surprisingly, it turned out that many of the concepts offered by other Balkan languages belonged to the same domains as the concepts chosen by the Serbian team. However, there were also concepts related to the plant and animal world, old arts and occupations, traditional music and dance, architecture, measuring units, etc.

In spite of the fact that Czech is not a language of the Balkans, the idea was to include it in the procedure in the same way as the other five languages. But since the Czech team did not specify its concepts on time, they could not take part in the procedure that followed step one. Step two thus established intersections of concepts between Bulgarian, Greek, Romanian, Serbian and Turkish. Some of the languages, including Serbian, added to their Balkan specific synsets a definition in English, besides the existing definition in their own language. The aim was to make the identification of common concepts easier. In addition to that, synsets offered by the Serbian team included examples of use extracted from the corpus of contemporary Serbian (Krstev and Vitas 2005).[4]

Out of 316 concepts selected by the Serbian team in phase one 109 were also proposed by at least one of the four remaining languages. Among these 109 concepts, the greatest number, 67 of them, were found also in the set of concepts proposed by the Bulgarian team. Greek followed with 37 common concepts, then Romanian with 29, and finally Turkish with 21.

After all the teams performed the task of identifying concepts they shared with other languages step two ended with the establishment of a set of 1562 different BSCs.

The only two concepts all five languages proposed in the first step as specific for their language were *кадаиф* and *алва* (Table 1), both of them representing condiments specific for the Balkan area. At the first glance it might look odd that other condiments from this area, even better known, such as *баклава* 'baklava' and *ратлук* 'Turkish delight' were omitted. The explanation lies in the fact that they did not satisfy the rule outlined in step one: they already existed in PWN.

Table 1.
The concepts *кадаиф* and *алва* in five Balkan languages

Bulgarian	Кадаиф	Халва
Greek	Καντα͡ΐφι	Χαλβάς
Romanian	Cataif	Halva
Serbian	Кадаиф	Алва
Turkish	Kadayıf	kağıt helva

There were 23 concepts proposed by four languages in step one, such as the ones shown in Table 2.

Table 2.
Some of the concepts common for four Balkan languages

	Minced meat	A family relation[5]	A dish made of intestines	A balcony in a religious building
Bulgarian	Кайма	Сват	Кавърма	Амвон
Greek		Συμπέθερος	Καβουρμάς	άμβωνας
Romanian	Carne_tocată	Cuscru		Amvon
Serbian	млевено	Пријатељ	Кавурма	
Turkish	Kıyma		Kavurma	Minber

Although they belong to different domains, most of them pertain to food and family relations. Some are in fact common for all five languages, but were not on all lists due to the fact that each team selected its concepts independently, and the omitted concepts were not given priority. For example, the concept *Балканијада* 'Balkan sport games' was not in the set of Turkish specific concepts, although it could have been. On the other hand, it is

understandable why concepts such as *стаљинизам* 'the period of Stalin' and *попадија* 'a wife of an orthodox priest' appeared in concepts specific for Bulgarian, Greek, Romanian or Serbian, but were not in the set of Turkish specific concepts.

It is interesting that all of the 23 concepts common for four languages were proposed in the set of Bulgarian specific concepts, and only seven of them were not offered for Serbian.

Out of 86 concepts shared by three languages, 45 appear in the set proposed by the Serbian team, e.g. *Други балкански рат* 'Second Balkan War,' *Втора Балканска война* in Bulgarian) and *Δεύτερος Βαλκανικός Πόλεμος* in Greek, or *Омладинска организација* 'socialist youth organization,' *комсомол* in Bulgarian and *UTC* in Romanian.

As the set of common concepts was determined following a procedure in which every team individually searched sets offered by other teams for concepts which conform with the concepts it proposed, some conflicts occurred. For example, it happened that one team stated that concept A in its language conforms to concept B in another language, whereas the team in charge for that language claimed that concept B is equivalent to concept C in the first language.

An even more complicated case is illustrated by the following example:

• the Bulgarian team claimed that *вуйчо* ⇔ Turkish *enişte*
• the Turkish team claimed that *enişte* ⇔ Serbian *метак*
• the Serbian team claimed that *метак* ⇔ Bulgarian *чичо*.

This conflict was resolved by making *вуйчо* ⇔ *enişte* ⇔ *метак* 'husband of one's aunt' equivalent. A small number of such conflicts remained unresolved, because partners could not agree how to relate mutually similar concepts.

3. Expanding Serbian Language Specific Concepts

In step three all teams expanded the set of their language specific concepts on basis of an analysis of concepts offered by other languages. The Serbian team started with the seven concepts offered for all four remaining Balkan languages, but omitted in the

initial set of Serbian specific concepts, as the most probable candidates. Indeed, six of them were recognized in Serbian, such as *шербе* 'sweet fluid,' which is also *шербе* in Bulgarian, *σερμπέτι* in Greek, and *şerbet* both in Romanian and Turkish.

As 45 out of 86 common concepts proposed for three languages were already included in SWN, the Serbian team continued with the analysis of the remaining 41. The majority of them, a total of 18, were those proposed by the Bulgarian, Greek and Romanian team. Ten of these 18 concepts were recognized in Serbian as well, such as *окрајак* 'the end of a bread loaf,' *крайщник* in Bulgarian, *γωνία* in Greek, and *coltuc* in Romanian. Out of 13 concepts offered by the Bulgarian, Greek and Turkish team, 9 were also recognized in Serbian, such as *зурле* 'a wind instrument,' *зурна* in Bulgarian, *ζουρνάς* in Greek, and *zurna* in Turkish. For Bulgarian, Romanian and Turkish there were five common concepts, four of which were recognized in Serbian, such as *шкембе* 'tripe soup,' *шкембе-чорба* in Bulgarian, *schembea* in Romanian, and *işkembe çorbası* in Turkish. The same number of concepts was proposed for Greek, Romanian and Turkish, and three of them were recognized in Serbian, such as *ока* 'unit of weight,' *οκά* in Greek, *oca* in Romanian, and *okka* in Turkish.

Following the same pattern the Serbian team turned to the 255 concepts common for two languages. As 54 of them were proposed in the first step as Serbian specific concepts, the team analyzed the remaining 201. The majority of concepts in this set were proposed for the Bulgarian-Greek language pair (72), and among them 47 were recognized in Serbian, one of them being *лазарка* 'girl praying for rain on St. Lazar's day,' *лазарка* in Bulgarian, and *Λαζαρίνες* in Greek. The smallest number of common concepts was proposed for the Romanian-Turkish language pair, only three, and only one of them was recognized in Serbian: *ћуфте* 'meat ball,' *chiftea* in Romanian and *çiğ köfte* in Turkish.

Finally the Serbian team analyzed the remaining 1,196 concepts proposed for only one language. However, this analysis was possible only to a limited extent. Namely, the Greek and Romanian partners have not included a definition in English for their concepts which made the comparison practically impossible. As for the 123 Bulgarian specific concepts, 48 of them were

recognized and included into SWN, among them, for example, *печење* 'roast meat,' *Васељенски патријарх* 'the Patriarh of Tsarigrad' and *фолк певачица* 'folk singer' (*чеверме*, *Вселенски патриарх*, and *фолк певица* in Bulgarian). Out of 202 Turkish specific concepts, 45 were recognized in Serbian such as *клањати* 'ritual prayer', *турбе* 'tomb of a famous Muslim', and *севап* 'good deed' (*namaz kılmak*, *türbe* and *sevap* in Turkish).

At the end, 223 new concepts were added to SWN, with 154 of them confirmed in the corpus of contemporary Serbian language and thus completed with extracted examples. Some of the concepts which could not be confirmed by the corpus can be considered as outdated, and are mostly based on Turkish etymology, such as *кајмакан* 'highest ranking administrative officer in a region,' and *рахле* 'low stand on which a book can be placed.' However, there were others, such as *зуџе* 'child's game played by hitting the partner with the palm in his palm placed in his armpit, from behind,' *маратор-салата* 'appetizer made of yoghurt, chopped cucumbers, garlic, mint and dill' and *ибришим* 'strong silk thread' confirmed by the (RMSMH 1967) and (Škaljić 1989) which still exist in the spoken language, but are not registered by the available corpus of written language. There was also a considerable number of concepts related to Islam which could not be confirmed by the corpus, for example *абдест* 'the act of cleaning one's body in line with the specified religious ritual' and *мувекит* 'an official who announces prayer times by observing the sky.'

Concepts common for several Balkan languages often have the same origin, mainly from Turkish. However, it should be noted that words of the same origin do not necessarily pertain to the same concept. An example is the Turkish specific concept *aktar*, *attar* 'shop where spices and herbs are sold.' One of the Serbian dictionaries (Škaljić 1989) contains the lemma *амар* (*atar)* with a related but different meaning "a person that sells medicines, a herb seller, a drug seller." The other dictionary (RMSMH 1967) does not contain a similar meaning for this lemma, nor does the lemma appear in the corpus of contemporary Serbian language.

4. Adding New Language Specific Concepts to SWN

After the process of introducing as many BCSs as possible into SWN according to the procedure described was completed, the Serbian team continued to expand SWN by adding language specific concepts denoting plant and animal species of Serbia, since many species well known in Serbia do not exist in PWN. However, a new species was added to SWN only in the case when the genus it belongs to already existed in PWN. If this was not the case, the addition of the concept was postponed, since it required specific knowledge to position the species in the frame of systemic division of the plant and animal world, which was not available at the moment. Addition of new species into SWN could be of considerable interest for other BalkaNet teams since many of these species are likely to be spread all over the Balkans. One of them is *врана* 'Corvus cornix,' *les kargasi* in Turkish, and *cioara griva* in Romanian.

Among other language specific features of Serbian pertaining to animals are concepts denoting young animals, which do not exist in PWN, such as *чавче, чавчић*, a young *чавка* 'jackdaw' or *магаре, пуле*, a young *магарац* 'donkey.' Closely related to them are concepts denoting the birth of a young animal, which are lexicalized by appropriate verbs, such as *ојарити се* 'give birth to a baby goat'. Similar concepts exist in Serbian for a number of various species, with their counterpart in PWN for only a few of them. In addition to that, concepts denoting the male and female representative of an animal species are often differently lexicalized in Serbian, as for instance *жаба* 'female toad' and *жабац* 'male toad.' There are no counterparts for such cases in PWN. Another related language specific feature is the supletive form of plural, that also exists for many animal species and which represents a group of them, like *јарад*.

5. Concluding Remarks

The establishment of Balkan specific concepts within BalkaNet demonstrated that besides concepts specific for certain domains, recognized as important and common in most of the languages that

developed their wordnets, important concepts specific for a certain language or a group of languages also exist. The procedure used for identifying BSCs within BalkaNet considerably enlarged the number of common concepts, and this number would be even bigger if the procedure included other Balkan languages.

There are further similarities among Balkan languages which could be used for the expansion of the set of BSCs. For example, Serbian has many concepts expressed by true reflexive verbs which would probably be recognized in the majority of Slavic languages and which do not exist in PWN, such as *волети се* 'to love each other.' Possessive adjectives derived from words which lexicalize noun concepts are another example. As a matter of fact, 80 possessive adjectives were proposed as Bulgarian specific concepts, and most of them, like *войнишки* 'that relates to a soldier and his service,' can be recognized in Serbian (*војнички*). Finally, concepts lexicalized by nouns resulting from gender motion are specific both for Bulgarian and Serbian, such as *омладинка* 'a girl member of the youth organization,' the female counterpart of *омладинац* 'a member of the youth organization.'

NOTES

[1] In Princeton WordNet the literal *transition* occurs in five senses, but not in the sense that is frequently used in the Balkans today: 'A period undergone by former socialist countries when the society and economics are adapted from socialism to capitalism.' For instance, the 25 million word corpus of contemporary Serbian (Krstev and Vitas 2005) records 1,116 instances of the word *транзиција* and almost all of them are in this sense, among them 196 within the compound *земља у транзицији* 'country in transition.'

[2] In the sense of 'embroidered cloth which usually hangs over the stove in the kitchen with humorous messages for the housewife.'

[3] *Ћевапчићи* 'meat fingers' is one of the most popular dishes in Serbian folk restaurants.

[4] However, 54 Serbian specific concepts had no confirmation in the corpus, like the adjective *деверов* 'belonging to *девер*,' as opposed to the noun *девер* 'the brother of one's husband,' from which the adjective is derived.

[5] In the sense 'father of one spouse to the father of the other spouse.'

REFERENCES

FELLBAUM, CHRISTIANE (ed.) 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

KRSTEV, CVETANA and DUŠKO VITAS. 2005. Corpus and Lexicon—Mutual In-completeness. Proceedings of the Corpus Linguistics Conference, ed. by Pernilla Danielsson and Martijn Wagenmakers, Birmingham. http://www.corpus.bham.ac.uk/PCLC/

RMSMH. 1967. Речник српскохрватскога књижевног језика (Dictionary of Serbo-Croat Literature Language). Нови Сад: Матица српска, Загреб: Матица хрватска.

ŠKALJIĆ, ABDULAH. 1989. Turcizmi u srpskohrvatskom jeziku (Words of Turkish origin in the Serbo-Croat Language). Sarajevo: Svjetlost.

TUFIŞ, DAN (ed.) 2004. Special Issue on BalkaNet Project, Romanian Journal on Information Science and Technology. Bucureşti: Publishing House of the Romanian Academy.

VOSSEN, PIEK (ed.) 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.

_____. 2004. Introduction to the Special Issue on the BalkaNet Project. In Tufiş 2004.

Cvetana Krstev
Faculty of Philology, University of Belgrade,
Studentski trg 3, 11000 Belgrade
[cvetana@poincare.matf.bg.ac.yu]

Ivan Obradović
Faculty of Geology and Mining, University of Belgrade,
Đušina, 11000 Belgrade
[ivano@rgf.bg.ac.yu]

Duško Vitas
Faculty of Mathematics, University of Belgrade,
Studentski trg 16, 11000 Belgrade
[vitas@poincare.matf.bg.ac.yu]