Developing Balkan specific concepts within *BalkaNet* - a multilingual database of semantic networks

Cvetana Krstev
Faculty of Philology, University of Belgrade, Studentski trg 3, 11000 Belgrade
Ivan Obradović
Faculty of Geology and Mining, University of Belgrade, Đušina, 11000 Belgrade
Duško Vitas
Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade

Abstract

The paper outlines the work involved in the construction of the multilingual semantic lexical database *WordNet* in the scope of the *Balkanet* project. The special focus of this paper are concepts specific for the Balkans, more precisely those that were not included in the Princeton *Wordnet* for English that has become the pillar for the construction of *Wordnets* for many languages. The methodology for the selection and acquisition of these concepts for Balkan languages, as well as the establishment of the potential equivalents is presented, as well as the encountered problems in its application. The domain and language characteristics of the established concepts are discussed. Special emphasize is given to the contribution of the Serbian specific concepts to this subset.

Introduction

The main goal of the Princeton WordNet (PWN), or simply WordNet, developed by George Miller and his associates was to serve as a sort of a mental lexicon in the scope of psycholinguistic research projects (Fellbaum 1998). PWN is a set of approximately 100.000 concepts interconnected by semantic relations to form a semantic network, where a concept denotes an abstract set of members grouped on basis of their common properties. Each concept in PWN is represented by a set of English synonyms which form the synset for this concept. A new dimension to the WordNet project was added by EuroWordNet (Vossen 1998) a project which introduced multilingualism into the semantic network. Namely, vocabularies for seven European languages have first been organized in a manner similar to PWN, and then mutually related via the so called Inter-Lingual-Index, or ILI.

The aim of BalkaNet, a project financed by the European Commission from September 2001 until August 2004 (Tufiş, 2004), was to develop aligned semantic networks for several Balkan languages, namely Bulgarian, Greek, Romanian, Serbian and Turkish, as well as to extend the existing network for Czech, initially developed within the EuroWordNet project. The consortium encompassed thirteen scientific and research institutions from Bulgaria, Greece, Romania, Serbia, Turkey, France, Netherlands and Czech Republic. Six teams were formed, one for each language. Besides the development of modern lexical resources for Balkan languages which would enable a new approach to the information originating from Balkan languages, the aim of BalkaNet was to extend the set of semantic networks established within the EuroWordNet project by adding networks for Balkan languages. One of the purposes of this extension was to reinforce the cooperation of Balkan countries with countries of the European Union.

The key activities within BalkaNet, that is, the development of WordNet type networks for individual Balkan languages and their connection with EuroWordNet, have been planned and executed synchronously. Namely, the monolingual networks were built on basis of commonly agreed sets with a total of 8.516 concepts already present in PWN. Beyond these sets, the network for each language has been developed independently, but within the framework set by PWN. Such an approach to the development of the network 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 offered confirmations of 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 classification of documents on a network or multilingual search. The existence of a multilingual database with mutually aligned concepts is thus of crucial importance.

Defining Balkan specific concepts

In order to overcome some of the problems encountered, all 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 all 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 is:

- 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'), or
- 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 Balkan specific concept has been rejected at the consortium level based on the conclusion that such a narrow determination would not be useful. Although there were supporters of the idea that the set of Balkan specific concepts 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.

It was agreed that the procedure of establishment of Balkan specific concepts defined in such a way should be executed in the following four steps:

1) Each team prepared a list of concepts specific for its language, cautiously verifying that the chosen concepts do not exist in PWN. Thus, for example, *баклава* '*baklava*', a natural candidate for a BSC, has not been included, since it already exists in PWN.
2) Each team compared its list prepared in step one with lists of concepts offered for other languages with the aim of linking its concepts with the similar concepts in other Balkan languages. Thus a multilingual core of BSCs was formed.
3) Each team inspected concepts offered by other teams for those recognizable in its language as well, and added them to its WordNet.
4) Finally, each team added new concepts to its WordNet.

In the first step 316 concepts (259 nouns, 9 verbs and 48 adjectives) which do not exist in PWN have been defined for Serbian and included in the Serbian WordNet (SWN). The majority of these concepts relate to food (*ајвар* 'pepper salad'), family relations (*јетрва* 'the wife of one's husband's brother'), society - mainly socialist heritage and transition (*ударник* 'a distinguished worker'), household (*кухарица* '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, (e.g. *ћевабџијин* 'belonging to *ћевабџија*' from *ћевабџија* 'one who produces and sells *ћевапчићи*'), with the verbs following a similar pattern (e.g. *партизановати* 'act as a *партизан*' from *партизан* 'partisan').

At the same time and independently from each other, the other teams described concepts specific for their languages. Thus 336 concepts were specified for Bulgarian, 309 for Greek, 545 for Romanian, 332 for Turkish and 226 for Czech. Many of the concepts offered for other Balkan languages belong to the same domains as the Serbian specific concepts, but there were also concepts related, to the plant and animal world, old arts and occupations, traditional music and dance, architecture, measuring units, etc.

The next step was dedicated to the establishment of intersections of concepts between languages. However, as the Czech team specified its concepts too late, they could not be compared with the concepts of other teams. Besides the concept definition that is the integral part of every synset, some of the languages, including Serbian, added a definition in

[1] In Princeton *WordNet*-y the literal *transition* occurs in five senses, but not in the sense that is frequently used in Balkan today 'A period undergone by former socialist countries when the society and economics are adapted from socialism to capitalism'. For instance, 25 million corpus of contemporary Serbian (Krstev and Vitas, 2005) records 1116 of the word *транзиција* and almost all of them are in thi sense, among them 196 in 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".

English to their Balkan specific synsets, in order to make the identification of common concepts easier. In addition to that, Serbian synsets include examples of use extracted from the corpus of contemporary Serbian[3] (Krstev and Vitas, 2005).

Out of 316 concepts offered for Serbian 109 of them were identified in at least one other language, whereas 199 of them were unique. The greatest number of common concepts was found among Bulgarian specific concepts, 67 of them. The same task was performed by other participants and this procedure ended with the establishment of a set of 1562 different Balkan specific concepts. Many of the concepts offered for other Balkan languages belong to the same domains as the Serbian specific concepts, but there were also concepts related, for example, to the plant and animal world, old arts and occupations, traditional music and dance, architecture, measuring units, etc.

The only two concepts offered for all five languages (Czech excluded) were *кадаиф* and *алва* (Table 1). Given that both of them represent a condiment, it might look odd that other condiments from the Balkan area, even better known, were not recognized. However, the best known of them: *баклава* 'baklava' and *ратлук* 'Turkish delight' were already in PWN.

Bulgarian	кадаиф	халва
Greek	κανταΐφι	χαλβάς
Romanian	cataif	halva
Serbian	кадаиф	алва
Turkish	kadayıf	kağıt helva

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

The 23 concepts which were common for four languages belong to different domains, but most of them pertain to food and family relations. Some of them (Table 2) are in fact common for all five languages, but were not on all lists due to the independent selection of concepts. 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 *сталъинизам* 'the period of Stalin' and *попадија* 'a wife of an orthodox *поп*' appear in specific concepts for Bulgarian, Greek, Romanian or Serbian, but were not in the set of Turkish specific concepts. All 23 concepts were proposed in the set of Bulgarian specific concepts, and only 7 of them were not offered for Serbian.

	Minced meat	**A family relation**	**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 месо	Пријатељ[4]	кавурма	
Turkish	Kıyma		kavurma	minber

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

Out of 86 concepts common for three languages, 45 appear in the set of Serbian specific concepts, e.g. *Други балкански рат* "The Second Balkan War', Втора Балканска война (Bulgarian), Δεύτερος Βαλκανικός Πόλεμος (Greek) and *Омладинска организација* 'The youth organization in SFRY', комсомол (Bulgarian), UTC (Romanian).

As the set of common concepts was determined following a procedure in which every team searched sets offered for other languages for concepts which conform with its concepts, some conflicts occurred. For example, it happened that one team stated that concept A in its language conforms to concept B of 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:

[3] 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', the adjective is derived from.

[4] In the sense "father of one spouse to the father of the other spouse ".

- 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.

Enlarging the set of Serbian language specific concepts

In the next step all teams, independently from one another, enlarged the set of their language specific concepts based on an analysis of concepts offered for other languages. For Serbian this step started with the analysis of the seven concepts offered for all other Balkan languages as the most probable candidates. Indeed, six of then were recognized in Serbian, such as *шербе* 'sweet fluid' σερμπέτι (Greek), şerbet (Turkish).

As 45 out of 86 common concepts proposed for three languages have been included in SWN in the first step, the remaining 41 were analyzed. The majority of them, a total of 18, were those proposed for Bulgarian, Greek and Romanian, and ten of these were recognized in Serbian as well, such as *окрајак* 'the end of a bread loaf', крайщник (Bulgarian), γωνία (Greek), coltuc (Romanian). Out of 13 concepts offered for Bulgarian, Greek and Turkish, 9 were recognized in Serbian, such as *зурле* 'a wind instrument', зурна (Bulgarian), ζουρνάς (Greek), zurna (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 (Bulgarian), schembea (Romanian), işkembe çorbası (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 (Greek), oca (Romanian), okka (Turkish).

Out of 255 concepts common for two languages, 54 were proposed in the first step as Serbian specific concepts. The majority of other concepts in this set have been proposed for the Bulgarian-Greek language pair (72), and among them was also the greatest number of those recognized in Serbian (47), one of them being *лазарка* 'girl praying for rain on St. Lazar's day', лазарка (Bulgarian), Λαζαρίνες (Greek). The smallest number of common concepts has been proposed for the Romanian-Turkish language pair, only three, and only one of them was recognized in Serbian: *ћуфте* 'meat ball', chiftea (Romanian), çiğ köfte (Turkish).

Finally the remaining 1196 concepts proposed for only one language has been analyzed, however, only partly. Namely, the Greek and Romanian partners have not included a definition in English for their concepts which made the comparison impossible. As for the 123 Bulgarian specific concepts, 48 of them have been recognized and included into SWN, among them, for example, *печење* 'roast meat', *Васељенски патријарх* 'he Patriarh of Tsarigrad' and *фолк певачица* 'folk singer' (чеверме, Вселенски патриарх, and фолк певица in Bulgarian). Out of 202 Turkish specific concepts, 45 of them were recognized in Serbian such as *клањати* 'ritual prayer', *турбе* 'tomb of a famous Muslim', and *севан* (namaz kılmak, türbe, and 'good deed' in Turkish).

In this final step 223 new concepts have been added to the SWN, with 154 of them confirmed in the corpus of contemporary Serbian language and thus completed with extracted examples. However, there were others, such as *зуце* ''child's game played by hitting the partner with the palm in his palm placed in his armpit, from behind', *таратор-салата* 'appetizer made of yoghurt, chopped cucumbers, garlic, mint and dill', and *ибришим* 'strong silk thread' confirmed by the (RMSMH, 1967) and (Škaljić, 1989) and still existing in the spoken language, but the corpus of written language does not register them. It should also be noted that a great number of concepts which could not be confirmed by the corpus can be considered as outdated, and based on Turkish etymology: *кајмакан* 'highest ranking administrative officer in a region', and *рахле* 'low stand on which a book can be placed'. It should also be noted that a considerable number of concepts related to Islam could not have 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, words of the same origin do not necessarily denote 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 "the person that sells medicines, a herb seller, a drug seller", whereas the other (RMSMH, 1967) does not contain a similar meaning for this lemma, nor does it appear in the corpus of contemporary Serbian language.

Further expansion of SWN

The SWN has additionally been expanded by language specific concepts denoting plant and animal species of Serbia, since many species well known in Serbia do not exist in PWN. However, this was done only in the case when the genus they belong to already existed in PWN. When this was not the case, the addition of the concept was postponed, since it requires precise positioning in the frame of the systemic division of the plant and animal world. The addition of the missing species into SWN can 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.
Even a single-sided analysis of all proposed concepts has considerably enlarged the number of common concepts. It is to be expected that this number will be considerably bigger when a similar analysis is performed for other Balkan languages included.

Further similarities among language specific concepts could also be detected. For instance, concepts lexicalized by nouns resulting from gender motion, 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'.

Among many other language specific features there are concepts related to young animals, which do not exist in PWN, such as *чавче, чавчић*, a young *чавка* 'jackdaw' or *магаре, пуле*, a young *магарац* 'donkey'. Related to these concepts are concepts denoting the birth of a young animal, lexicalized by appropriate verbs. Such concepts exist in Serbian for a number of various species, with their counterpart in PWN for only a few of them. An example is *ојагити се* defined as "give birth to a baby goat". In Serbian, for many animal species concepts are lexicalized that denote the male or female representative, for instance *жабац* 'male toad' for which there is no counterpart in PWN. A 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*.

Concluding remarks

The establishment of Balkan specific concepts within the development of the semantic lexical database *Balkanet* demonstrated that besides domain specific concepts, which the participants were mainly occupied by in the initial phase of the project, there is great number of those that are specific for one or more languages. The procedure used for identifying BSCs within Balkanet considerably enlarged the number of common concepts. It is to be expected that this number would be even bigger if the procedure included other Balkan languages.

The set of BSCs could be further expanded by other types of language specific concepts. For example, concepts expressed by true reflexive verbs which do not exist in PWN and which would probably be recognized in all Slavic languages included. An example is *волети се* 'to love each other', which does not exist as a PWN concept. Another type are possessive adjectives derived from words which lexicalize both language specific noun concepts and noun concepts that appear in PWN. It should be noted that 80 of them 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* and Greek as well.

References

Fellbaum, C. ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Krstev, C. & D. Vitas. 2005. Corpus and Lexicon - Mutual Incompleteness. In P. Danielsson and M. Wagenmakers eds., *Proceedings of the Corpus Linguistics Conference*. 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, A. 1989. *Turcizmi u srpskohrvatskom jeziku (Words of Turkish origin in the Serbo-Croat Language)*. Sarajevo: Svjetlost.

Tufiş, D. ed. 2004. *Special Issue on BalkaNet Project*, Romanian Journal on Information Science and Technology. Bucureşti: Publishing house of the Romanian academy.

Vossen, P. ed. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

Vossen, P. 2004. Introduction to the Special Issue on the BalkaNet Project. In Tufiş 2004.