Resources for Processing Bulgarian and Serbian – a brief overview of Completeness, Compatibility, and Similarities

Svetla Koeva, Cvetana Krstev, Ivan Obradović, Duško Vitas

Department of Computational Linguistics – IBL, BAS
52 Shipchenski prohod, Bl. 17
Sofia 1113
svetla@ibl.bas.bg

Faculty of Philology
University of Belgrade
Studentski trg 3
11000 Belgrade
cvetana@matf.bg.ac.yu

Faculty of Geology and Mining, U. of Belgrade
Đušina 7
11000 Belgrade
ivano@rgf.bg.ac.yu

Faculty of Mathematics
University of Belgrade
Studentski trg 16
11000 Belgrade
vitas@matf.bg.ac.yu

Abstract

Some important and extensive language resources have been developed for Bulgarian and Serbian that have similar theoretical background and structure. Some of them were developed as a part of a concerted action (wordnet), others were developed independently. Brief overview of these resources is presented in this paper, with emphasis on similarities and differences in the information presented in them. Special attention is given to similar problems encountered in the course of the development.

1. Introduction

Bulgarian and Serbian as Slavonic languages show similarities in their lexicons and grammatical structures. At present, some equivalent language resources were developed for both languages, moreover the formal approaches for the organization of those recourses are very similar. The goal of this paper is briefly to present these similarities while describing the integration between electronic dictionaries and lexical-semantic data bases (wordnets) of both languages.

The Bulgarian and Serbian wordnets have been initially developed in the framework of the project ***BalkaNet – a multilingual semantic network for the Balkan Languages*** which has been aimed at the creation of a semantic and lexical network of the Balkan languages [Stamou, 2002] with a view to their integration in the global WordNet [Fellbaum, 1998; Miller, 1990] — an extensive network of synonymous sets and the semantic relations existing between them in different languages, enabling cross-references between equivalent sets of words with the same meaning [Vossen, 1999].

The common origin of Bulgarian and Serbian, the equivalent types of existing electronic resources and application approaches used offer not only a very good basis for comparative research but furthermore presupposes the successful implementation in such different application areas as cross-lingual information and knowledge management, cross-lingual content management and text data mining, cross-lingual information retrieval and information extraction, multilingual summarization, multilingual language generation etc.

2. Electronic dictionaries

2.1 Bulgarian Grammatical Dictionary

The grammatical information included in the Bulgarian Grammatical Dictionary (BGD) is divided into three types [Koeva, 1998]: category information that describes lemmas and indicates the words clustering into grammatical classes (Noun, Verb, Adjective, Pronoun, Numeral, and Other); paradigmatic information that also characterizes lemmas and shows the grouping of words into grammatical subclasses, i.e. — Personal, Transitive, Perfective for verbs, Common, Proper for nouns, etc.; and grammatical information that determines the formation of word forms and shows the classification of words into grammatical types according to their inflection, conjugation, sound and accent alternations, etc.

The BGD is a list of lemmas where each entry is associated with a label [Koeva, 2004a]. The label itself represents the grammatical class and subclass to which the respective lemma belongs and contains a unique number that shows the grammatical type. All words in the language that belong to the same grammatical class, subclass and have an identical set of endings and sound / stress alternations are associated with one and the same label. Each label is connected with the corresponding formal description of endings and alternations. The inflectional engine used is equivalent to a stack automaton. Despite the existence of some

CATEGORY	BULGARIAN	SERBIAN
Gender		masculine, feminine, neutral
Number	singular, plural, counting form	singular, plural, paucal
Case	vocative	nominative, genitive, dative, accusative, vocative, instrumental and locative
Definiteness	definite, indefinite, definite - full form, definite - short form	
Animateness		animate, inanimate

Table 1. Grammatical features of nouns

differences in the format, the BGD represents a kind of DELAS dictionary [Courtois, 1990; Courtois at al, 1990; Silberztein, 1990], and it is compiled into a Finite-State Transducer.

2.2. Serbian Morphological Dictionary

Electronic dictionaries of Serbian consist of morphological dictionaries of general lexica, dictionaries of proper names, and the Serbian wordnet. The system of e-dictionaries of simple words in Serbian has been developed according to the LADL model and described, with other types of dictionaries, in [Vitas & al., 2003]. Following this model the system is based on a dictionary of lemmas named DELAS. A dictionary of all inflectional forms, named DELAF, is automatically generated on basis of morphological information attached to lemmas in DELAS. The most important piece of information accompanying a DELAS lemma is the inflectional class it belongs to, which enables the generation of all inflected forms of a lemma with accompanying grammatical information. The information on the inflectional class is expressed by a code, e.g. N600 or V651. The information attached to a lemma in the DELAS dictionary relates to all forms of that lemma, whereas morphological information attached to the inflected form in the DELAF dictionary is characteristic of that form only.

3. Morphological information in electronic dictionaries of Bulgarian and Serbian

With respect to PoS (parts of speech), the Princeton Wordnet (PWN) and other wordnets that use PWN as a model consist of nouns, verbs, adjectives and adverbs.

3.1. Nouns

Nouns in Bulgarian and Serbian are characterized by their inflectional categories (Table 1).

Bulgarian nouns are divided into grammatical subclasses with respect to their type (Common, Proper, Singularia tantum, Pluralia tantum) and Gender. The category Gender with Bulgarian nouns is a lexical-semantic category, which means that a given noun does not possess different word forms expressing masculine, feminine and neuter, although the noun lemmas can be grammatically classified into the three classes: *стол* (chair) - masculine, *маса* (table) – feminine, and *куче* (dog) - neuter.

The category Case has lost its morphological realization in the system of Bulgarian nouns and only vocative against nominative is kept with some proper and common nouns (in masculine and feminine) denoting persons. Some concrete nouns also allow potential generation of vocative in metaphorical usage.

The category Definiteness is realized by means of indefinite and definite forms that add a definite morpheme at end of the word. A special feature of Bulgarian is the existence of two definite morphemes for masculine distinguishing the syntactic functions of subject from others.

Bulgarian masculine non-animate nouns after counting numerals and quantifiers are used in plural in a counting form – *пет учебника* (five textbooks), *десет бора* (ten pine-trees).

In Serbian, nouns are morphologically realized in seven cases. The category Gender is in Serbian an inflectional category: for instance *papa* (pope) is masculine but its plural form *pape* is feminine. Besides two main categories for number, singular and plural, Serbian nouns also have the so called “paucal” form which represents a synthetic category of number and gender that is used with small numbers (two, three four): *jedan lep zec* (one pretty rabbit), *dva lepa zeca* (two pretty rabbits), *pet lepih zečeva* (five pretty rabbits). Animateness is also the inflectional category for masculine gender nouns: the form of the accusative

CATEGORY	BULGARIAN	SERBIAN
Person	first, second, third	first, second, third
Number	singular, plural	singular, plural
Tense	present, aorist, imperfect	present, aorist, imperfect, future
Mood	indicative, imperative	infinitive, imperative
Participles	present active, aorist active, imperfect active, past passive	past active, past passive
Voice	active, passive	active, passive
Definiteness	definite, indefinite, definite – full form, definite - short form	
Gender	masculine, feminine, neuter	masculine, feminine, neuter
Gerund	past active	present active, past active

Table 2. Grammatical features of verbs

case is equal to the genitive case for the animate nouns and to the nominative case for the inanimate nouns.

Noun lemmas in the Serbian DELAS dictionary are marked with markers which sometimes determine the noun in a more precise manner. For example, pluralia tantum are marked with the marker +PT as in *pantalone* denoting the concept lexicalized in PWN as {*trousers:1, pants:1*} and in the Bulgarian wordnet as {*панталони*:1, *панталон*:1}. The markers +MG +FG are used to mark the natural male and female gender (or sex) which does not necessarily match the grammatical gender and which is important for agreement. This is the case, for example, with the noun *izbeglica* (refugee), which denotes persons of both male and female sex. This noun is inflected as a noun of feminine gender, agrees with the adjective as a noun of feminine or masculine gender in singular (*za svakog* (m) *izbeglicu* (f) – for every refugee) and as a noun of feminine gender in plural, and can agree with the relative pronoun in plural both as a noun of feminine gender (*Izbeglice* (f) *koje* (f) *su juče stigle* (f) *su izjavile* (f)… – The refugees that arrived yesterday said…) and as a noun of masculine gender (*UNHCR će pružiti pomoć za izbeglice* (f) *koji* (m) *žele da se integrišu u lokalnu sredinu* – UNHCR will provide help for refugees that want to integrate into the local society). Finally, the marker +Pl marks a noun in singular which denotes a natural plural: *braća* (brothers) is inflected as a noun of feminine gender in singular and agrees as noun both with singular and plural: *Njena* (s) *braća* (s) *su* (p) *dolazila* (s) *svaki dan* (her brothers came every day).

3.2. Verbs

Verbs in Bulgarian and Serbian are characterized by inflectional categories and their values (see Table 2). Bulgarian Verbs are classified in subclasses with respect to Transitivity (transitive and intransitive), Perfectiveness (perfective and imperfective), and Personality (personal, third personal and impersonal), while the Serbian verbs are classified according to the first two features.

Verb lemmas in Serbian are characterized by the following markers: for aspect imperfective +Imperf and perfective +Perf, for reflexiveness reflexive +Ref and irreflexive +Iref, and for transitivity transitive +Tr and intransitive +It.

Many verbs in the two languages can be both imperfective and perfective, such as *адресирам* and *adresirati* which denote the concept lexicalized in PWN as {*address*:3, *direct*:12}. Many formally equal verbs can express both reflexive and irreflexive meaning, such as {*topiti:*1a} lexicalized in PWN as {*melt*:1, *run*:39, *melt down*:1} and in the Bulgarian wordnet as {*топя*:1, *стопявам*:1, *стопя*:1, *разтопявам*:1, *разтапям*:1, *разтопя*:1} and *topiti se*, lexicalized in PWN as {*dissolve*:9, *thaw*:1, *unfreeze*:1, *unthaw*:1, *dethaw*:1, *melt*:2} and in the Bulgarian wordnet as {*топя се*:1, *стопявам се*:1, *стопя се*:1, *разтопявам се*:1, *разтапям се*:1, *разтопя се*:1}. Lexical reflexivity in both languages is expressed by the lexical particle *se* (and *si* for Bulgarian). Cognate verbs can also express either transitive or intransitive meaning, such as {*svirati:*1b} denoting the concept lexicalized in PWN as {*play*:3} and in the Bulgarian wordnet as {*свиря*:3} as intransitive verb (*The band played all night long*), or {*svirati*:1a} denoting the concept lexicalized in PWN as {*play*:7} and in the Bulgarian wordnet as {*свиря*:1} as transitive verb (*He plays the flute*). Synsets that contain the same verbs, in one case as reflexive and in the other as irreflexive are often

CATEGORY	BULGARIAN	SERBIAN
Gender	masculine, feminine, neutral	masculine, feminine, neutral
Number	singular, plural	singular, plural, paucal
Case		nominative, genitive, dative, accusative, vocative, instrumental, and locative
Definiteness	definite, indefinite, definite - full form, definite - short form	definite, indefinite
Comparison	positive , comparative, superlative	positive, comparative, superlative
Animateness		animate, inanimate

Table 3. Grammatical categories with adjectives

linked by the *cause / caused* relation. The transitive / intransitive forms have to have separate meanings in PWN. This is not the case with the aspect – perfective verbs which are not generated by prefixing should be in the same synset as the imperfective verb.

The Bulgarian imperative has two declined forms - 2nd person singular and plural, in comparison with Serbian where three declined forms: 2nd person singular, 1st and 2nd person plural, are realized.

Bulgarian participles are specified for aspect and are declined according to number, gender and definiteness. Serbian participles are specified for aspect and are declined according to number: singular or plural, and gender: masculine, feminine, or neutral.

Participles in both languages are used to form compound tenses, both in the active and passive voice: perfect, pluperfect, future (past and perfect), and conditional.

Infinitive in Serbian and Gerunds in both languages are indeclinable.

3.3. Adjectives

The categories realized with adjectives in Bulgarian and Serbian are similar as well. The main differences are observed with the categories Case and Animateness. Adjectives are characterized by their morphological categories and their members (shown in Table 3).

3.4. Adverbs

Adverbs in traditional Bulgarian and Serbian grammars are considered indeclinable word types, although for many of them comparison exists: for example, *бързо, brzo* (rapidly), *по-бързо, brže* (more rapidly) and *най-бързо, najbrže* (the most rapidly). These are usually treated as separate lemmas but a paradigmatic analysis in the scope of the morphological category Comparison is also acceptable. There is another level of comparison for adjectives and adverbs in Serbian which is realized by the prefix *po-* and superlative: *ponajbrže*, which relativizes the superlative and denotes, in this case, the fastest way among the slow ways.

4. Language resources integration

4.1. Bulgarian resources

There are three large Bulgarian resources: Bulgarian WordNet (BulNet) which covers approximately one third of the general Bulgarian lexicon [Koeva & al., 2004], BGD - encoding lemmas and corresponding inflection types and Bulgarian Frame Lexicon - encoding the arguments of the verbs and their semantic features [Koeva, 2004c]. The combination of these resources results in their mutual enhancement, their expansion and reliable validation.

The Bulgarian WordNet models nouns, verbs, adjectives, and (occasionally) adverbs, and contains already 24 405 synonymous sets (towards 1.09.2005), where 51 584 literals have been included (the ratio is 2,11). Following the standards accepted in the BalkaNet project the structure of the Bulgarian and Serbian data bases is organized in an XML file. Every synset encodes the equivalence relation between several literals (at least one has to be present), having a unique meaning (specified in the SENSE tag value), belonging to one and the same part of speech (specified in the POS tag value), and expressing the same lexical meaning (defined in the DEF tag value). Each synset is related to the corresponding synset in the English Wordet2.0 via its identification number ID. The common synsets in the Balkan languages are encoded in the tag Base Concepts -- BCS. There has to be at least one language-internal relation (there could be more) between a synset and another synset in the monolingual data base. There could also be several optional tags encoding usage, some stylistic, morphological or syntactical features, a stamp marking the person who worked out the particular synset, as well as the last time it was edited.

In order to merge the language data existing in BulNet and BGD it was decided to assign an additional grammatical note to each literal thus linking it with the BGD lemma's label [Koeva, 2004b]. All labels for BGD entry forms that are found in the BulNet have been entered as values of the LNOTE grammatical tag in the XML format. Most of the literals which were not recognized are either specialized terms that have no place in a grammatical dictionary of the common lexis (often written in Latin) or compounds. The contradictory cases where two or more labels were associated with one and the same literal were solved manually.

The grammatical specifications used in the Bulgarian Frame Lexicon are identical with those in BGD and BulNet. Thus the Bulgarian WordNet is fully exportable with the syntactic information available from the Frame lexicon.

4.2. Serbian resources

The Serbian wordnet covers at present approximately one fifth of the Serbian general lexicon, but it is constantly being developed [Krstev & al., 2004a]. In the course of its development it has been enriched with information pertaining to inflexion of literals – simple words. A software tool specially designed for this purpose is used to enable automatic transfer of all information on the inflectional class of a literal from the morphological dictionary into the wordnet where it becomes the content of the <LNOTE> element for that literal (the <LNOTE> element is part of the content of the <LITERAL> element) [Krstev & al., 2004b]. The program allows the user to alter the automatically assigned class in cases when different choices are possible.

Inflections are of great importance for Serbian language, given the fact that the generation of inflective forms is not straightforward. This can be best illustrated by the existence of a large number of homograph lemmas: for example, *deka* can be a synonym for a blanket, a unit of measurement (short for decagram) or a hypocoristic for grandfather[1]. In the first two cases the nouns are inanimate, and of feminine gender, with the same inflection – they belong to one and the same inflectional class. In the third case the noun is animate, of masculine gender in singular and feminine gender in plural form, and belongs to different inflective class.

4.3. Problems to be solved

Literals – simple words, can appear in the Bulgarian and Serbian wordnets which are not lemmas in the morphological dictionary. Such is the case with animal and plant species, which appear as nouns in plural – the singular denotes just one member of the species. For example, the Serbian wordnet contains the synset {*Felidae*:1, *porodica Felidae*:1, *mačke*:X}, where the value N603+Zool:p has been assigned to the <LNOTE> element for the last literal – which means that the literal belongs to the N603 inflective class (fleeting "a" appears in genitive plural), is marked as animal (+Zool), an is always used in plural (:p). The corresponding Bulgarian synset is {*Котки*:1, *семейство Котки*:1, *Фелиде*:1, *семейство Фелиде*:1, *Felidae*:1, *семейство Felidae*:1}, and the English is {*Felidae*:1, *family Felidae*:1} which belongs to the hierarchical branch that starts with {*group:1, grouping:1*}.

group:1, grouping:1
  biological group:1
    taxonomic group:1, taxonomic category:1, taxon:1
      family:6
        mammal family:1
          Felidae:1, family Felidae:1

On the other hand *mačka* (cat) from the synset {*životinja iz roda mačaka*:1, *mačka*:1b} (corresponding to the PWN synset {*feline*:1, *felid*:1}, and the Bulgarian synset {*фелид*:1}) which belongs to the hierarchical branch that starts with {*organism:1, being:2*} is in a holo_member relation with the former synset has N603+Zool as the content of the <LNOTE> element, which means that the noun can appear both in singular and plural. This synset belongs to the hierarchical tree branch:

organism:1, being:2
  animal:1, animate being:1, beast:1, …
    chordate:1
      vertebrate:1, craniate:1
        mammal:1
          placental:1, placental mammal:1, …
            carnivore:1
              feline:1, felid:1

Many literals in the Bulgarian and Serbian wordnets, as in other wordnets, are not simple words but compounds. There are 12 636 compound literals out of 51 584 in BulNet (24,49 %) and respectively 3 081

[1] All three lemmas are accentuated in a different manner, but that is not obvious from written text.

such literals out of 16 621 existing in Serbian WordNet (18,53 %). The majority of them fall in one of the following categories:

1.Adjective*-noun, for example {*konusni presek*:1, *kupasti presek*:1} (corresponding to {*conic section*:1, *conic*:1} in PWN and {*конично сечение*:1}) in Bulgarian wordnet, or {*konjska trka*:1} (corresponding to {*horse race*:1} in PWN and {*конно състезание*:2, *конно надбягване*:1} in the Bulgarian wordnet).

2. Noun phrases where the noun is supplemented with a prepositional phrase: for example, {*pobeda na poene*:1} (corresponding to {*decision*:3} in PWN and {*победа по точки*:1} in the Bulgarian wordnet), or {*daska za peglanje*:1} (corresponding to {*ironing board*:1} in PWN and {*дъска за гладене*:1} in the Bulgarian wordnet).

3. Coordinate noun phrases (just a few), such as *muž i žena* in {*bračni par*:1, *muž i žena*:1} (corresponding to {*marriage*:2, *married couple*:1, *man and wife*:1} in PWN and {*съпружеска двойка*:1, *съпрузи*:1, *семейна двойка*:1, *мъж и жена*:1} in the Bulgarian wordnet).

4. Verb phrase in which the verb is supplemented by a noun phrase, such as {*живея:3, водя живот*} corresponding to {*live:2*} *in PWN and* {*živeti život:1, voditi život:1*} in the Serbian WordNet.

5. A genitive phrase in Serbian: such as {*deljenje akcija*:1} (corresponding to {*split*:9, *stock split*:1, *split up*:1} in PWN and {*стоксплит*:1} in the Bulgarian wordnet), or {*izraz lica*:1} (corresponding to {*countenance*:1, *visage*:2}, in PWN and {*израз*:2, *изражение*:1} in the Bulgarian wordnet).

6. Noun-noun subordinate phrase in Serbian, which are the rarest: for example {*biljka penjačica*:1} (corresponding to {*vine*:1} in PWN and {*увивно растение*:1, *пълзящо растение*:1} in the Bulgarian wordnet),

Compounds have their own inflective rules: for example, in the second and fifth case only the head noun is inflected, whereas in the third and sixth case both nouns are inflected. In the fourth case only the verb is inflected in Serbian and Bulgarian (in this particular case). In the first case the noun is inflected and the adjective(s) agree with the noun. A precise description of this type of inflections remains to be elaborated in accordance with the solution proposed in [Savary, 2005]. This is why the <LNOTE> elements for compounds in Bulgarian and Serbian wordnets still remain empty. In Bulgarian and Serbian wordnets, as in PWN, there are a lot of Latin names for species that are uninflected in practice.

5. Mirroring of PWN concepts and structure to Bulgarian and Serbian

The BalkaNet project adopted the Princeton WordNet structure and concepts as the model for the development of wordnets for five Balkan languages and Czech. However, the development of these wordnets showed that mirroring PWN synsets and the relations among them to Balkan languages is neither the simplest nor the most appropriate solution. Its rationale could be found principally in the necessity of obtaining a coherent multilingual lexical database. The problems encountered were many. We will illustrate some of them with examples related to Serbian and Bulgarian.

The simplest problem was the absence of specific PWN concepts in Serbian and/or Bulgarian. An example is the PWN concept defined as “an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience” and lexicalized as synset {*plant*:4}. Although the synsets for this concept have been introduced in both the Serbian and the Bulgarian wordnet, the lexicalizations in Serbian {*glumac iz publike*:1} and Bulgarian {*подставен актьор*:1, *актьор в публиката*:1} in fact do not adequately represent the original PWN concept.

Conversely, the problem of absence of Serbian and/or Bulgarian concepts as well as concepts from other BalkaNet languages in PWN was also encountered. The solution for this problem was sought within the project in the introduction of the *language specific* and *Balkan specific* concepts. Initially, a set of concepts, not present in PWN, was defined for each language, with appropriate synsets and an English definition attached.

In this stage 316 Serbian specific concepts were defined: 259 nouns, 9 verbs and 47 adjectives. There were 336 concepts defined for Bulgarian, 309 for Greek, 545 for Romanian, 332 for Turkish and 226 for Czech. The English definition attached to the appropriate synsets enabled mutual comparison of language specific concepts, and extraction of concepts common for two or more languages, such as two oriental sweets common for Bulgarian, Greek, Romanian, Serbian and Turkish (Fig 1), defined in all five initial sets of language specific concepts for these languages, and nonexistent in PWN.

Every language specific concept became a Balkan specific concept. These concepts were incorporated into appropriate BalkaNet wordnets, and common concepts were linked via a BILI (BalkaNet ILI) index.

Bulgarian	кадаиф	Халва
Greek	κανταΐφι	Χαλβάς
Romanian	cataif	Halva
Serbian	кадаиф	Алва
Turkish	kadayıf	helva

Figure1. Two Balkan specific concepts common to five languages

The initial set of Balkan specific common concepts consisted mainly of concepts reflecting the cultural specifics of the Balkans (many of them pertaining to family relations, religion, socialist heritage etc.). Serbian wordnet presently contains 538 Balkan specific and 55 Serbian specific concepts, Bulgarian – 444 Balkan specific and 42 Bulgarian specific concepts.

There are other specific features of Bulgarian and Serbian that are of a linguistic nature and that disable the strict one-to-one mapping with PWN. For example, a very small number of possessive and relative adjectives can be found in PWN, whereas the initial set of language specific concepts for Bulgarian contained a number of relative adjectives, most of them having an equivalent in Serbian. For example, the relative adjective {*стоманен*:1} defined in Bulgarian as "който се отнася до стомана" (of or related to steel) has the Serbian equivalent {*čelični*:1} with exactly the same definition "koji se odnosi na čelik". Another example is {*войнишки*:1} defined in Bulgarian as "който се отнася до войник или войнишка служба" (of or related to a soldier and army service) which has the Serbian equivalent {*vojnički*:1} with practically the same definition "koji se odnosi na vojnika ili njegovu službu". Another group of concepts specific both for Bulgarian and Serbian (but also for some other BalkaNet languages) are lexicalized by nouns resulting from gender motion. Some of them were accepted as Balkan specific concepts. For example, {*omladinac*:1} defined as "član, pripadnik omladinske organizacije" (a member of the youth organization.) has its female gender counterpart lexicalized by a noun derived by gender motion {*omladinka*:1} defined as "devojka, član omladinske organizacije" (a girl, member of the youth organization.). Both concepts, also related by gender motion, exist in Bulgarian: {*комсомолец*:1} and {*комсомолка*:1}. For some concepts which exist in PWN, such as {*politician*:2, *politico*:1, *pol*:1, *political leader*:1}, which have their Serbian and Bulgarian equivalent in {*političar*:1} and {*политически лидер*:1}, there is no corresponding concept in PWN related to the female gender, whereas such a concept, again lexicalized by a noun derived by gender motion, exists in Serbian: {*političarka*:1}. In order to describe relations between concepts in the aforementioned cases, specific relations, more specific than the *derived* relation already existing in PWN, were introduced in the Serbian wordnet, namely: *derived-pos* and *derived-gender*. However, all these relations are in general inadequate, since they link synsets rather than literals, whereas the relation of derivation can only pertain to literals.

Among many other language specific features we mention here also concepts related to young animals, which do not exist in PWN, such as {*čavče*:1, *čavčić*:1}, a young *čavka* (jackdaw) or {*jare*:1, *jarence*:1, *kozlić*:1}, a young *koza* (goat). 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 {*ojariti se*:1} defined as "give birth to a goat". The same features are shown in Bulgarian although the equivalent examples are not yet included in BulNet.

A specific problem is posed by concepts lexicalized by nouns originating from regular derivation which does not alter either the PoS or the gender, such as diminutives and augmentatives [Vitas & Krstev, 2005]. There are several possible approaches to these nouns:

- treat them as denoting specific concepts and define appropriate synsets;
- include them in the synset with the noun they were derived from;
- omit their explicit mentioning, but rather let the flexion-derivation description encompass these phenomena as well.

The first approach is mandatory if the diminutive or augmentative acquires a special meaning: for example, the diminutive *glavica* from *glava* (head) is used in Serbian for the concept lexicalized in English as {*head cabbage*:1, *head cabbage plant*:1, *Brassica oleracea* *capitata*:1} whereas the augmentative *glasina* from *glas* (voice) is used for the concept lexicalized in English as {*rumor*:1, *rumour*:1, *hearsay*:1} and in Bulgarian as {*слух*:2, *мълва*:1, *клюка*:1} and defined as “gossip (usually a mixture of truth and untruth) passed”. On the other hand, if the third approach is accepted, the question arises whether it is possible to apply the same approach to other regular phenomena (gender motion and possessive adjectives)?

6. Conclusions

For both languages the importance of including inflectional information into the wordnet has been recognized and, consequently, it was added in the wordnets for the respective languages. However, a lot of work still remains to be done, particularly for the inflectional description of compound words. The first results obtained by the comparison of the extensive and powerful resources already developed promise their possible successful usage in many NLP applications.

References

[Courtois, 1990] Courtois, B. (1990). Le dictionnaire DELAS. in *Dictionnaires électroniques du français*, Langue française n° 87 (pp. 11-22). Larousse: Paris.

[Courtois at al., 1990] Courtois B., Silberztein, M. Eds (1990). *Dictionnaires électroniques du français*, Langue française n° 87 (127 pages). Larousse: Paris.

[Fellbaum 1998] Fellbaum, C. (ed.). WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press, 1998. [Koeva, 1998] S. Koeva *Bulgarian Grammar Dictionary. Description of the linguistic data organization concept* in: Bulgarian language, 1998, 6, 49-58.

[Koeva at al., 2004] S. Koeva, T. Tinchev and S. Mihov *Bulgarian Wordnet-Structure and Validation* in: Romanian Journal of Information Science and Technology, Volume 7, No. 1-2, 2004: 61-78.

[Koeva, 2004a] S. Koeva M*odern language technologies – applications and perspectives,* in: Lows of/for language, Hejzal, Sofia, 2004, 111- 157

[Koeva, 2004b] S. Koeva *Validating Bulgarian WordNet using grammatical information* in: Proceedings from Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, Göteborg University, 2004, 80-*82*

[Koeva, 2004c] Koeva S., *Theoretical model for a formal representation of syntactic frames*, Scripta and e-Scripta. Vol.2 , Sofia,2004: 9-26.

[Krstev at al., 2004a] C. Krstev, G. Pavlović-Lažetić, D. Vitas, I. Obradović (2004a) . Stamou, K. Oflazer, K. Pala, D. Christodoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, M. Grigoriadou *Using Textual and Lexical Resources in Developing Serbian Wordnet* in: Romanian Journal of Information Science and Technology, Volume 7, Numbers 1-2, 2004, pp. 147-161.

[Krstev at al., 2004b] C. Krstev, D. Vitas, R. Stankovic, I. Obradovic, G. Pavlovic-Lazetic, (2004b) Combining Heterogeneous Lexical Resources in Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisabon, Portugal, May 2004, vol. 4, pp. 1103-1106, ARTIPOL - Artes Tipograficas, Lda, Portugal.

[Miller, 1990] Miller G. A. Introduction to WordNet: An On-Line Lexical Database. In ``International Journal of Lexicography", Miller G.A., Beckwidth R., Fellbaum C., Gross D., Miller K.J. Vol. 3, No. 4, 1990, 235--244. [Savary, 2005] A. Savary, (2005) Towards a Formalism for the Computational Morphology of Multi-Word Units in Proceedings of 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland, ed. Zygmunt Vetulani, pp. 305-309, Wydawnictwo Poznanskie Sp. z o.o., Poznan.

[Silberztein, 1990] Silberztein, M. Le dictionnaire DELAC. In *Dictionnaires électroniques du français*, Langue française n° 87 (pp. 73-83). Larousse: Paris.

[Stamou, 2002] Stamou S., K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufis, S. Koeva, G. Totkov, D. Dutoit, M. Grigoriadou, BALKANET: A Multilingual Semantic Network for the Balkan Languages, Proceedings of the International Wordnet Conference, Mysore, India, 21-25 January 2002, 12-14.

[Vitas at al., 2003] D. Vitas, C. Krstev, I. Obradović, Lj. Popović, G. Pavlović-Lažetić (2003) *An Overview of Resources and Basic Tools for the Processing of Serbian Written Texts* in: Workshop on Balkan Language Resources and Tools, Novembar 21, Thessaloniki, Greece.

[Vitas & Krstev, 2005] Duško Vitas, Cvetana Krstev (2005) *Derivational Morphology in an E-Dictionary of Serbian* in Proceedings of 2nd Language & Technology Conference, April 21-23, 2005, Poznań, Poland, ed. Zygmunt Vetulani, pp. 139-143, Wydawnictwo Poznańskie Sp. z o.o., Poznań, 2005.

[Vossen, 1999] Vossen P. (ed.) EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer Academic Publishers, Dordrecht. 1999.