Morpho-semantic Relations in Wordnet – a Case Study for two Slavic Languages

Svetla Koeva[1], Cvetana Krstev[2], Duško Vitas[3]

[1] Department of Computational Linguistics, Institute of Bulgarian, 52 Shipchenski prohod, 1113 Sofia, Bulgaria
svetla@ibl.bas.bg
[2] Faculty of Philology, University of Belgrade, Studentski trg 3, 11000 Belgrade, Serbia
[3] Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
{cvetana, vitas}@matf.bg.ac.yu

Abstract. In this paper we present the problem of representing the morpho-syntactic relations in wordnets, especially the problems that arise when wordnets for languages that differ significantly from English are being developed on the basis of the Princeton wordnet, which is the case for Bulgarian and Serbian. We present the derivational characteristics of these two languages, how these characteristics are presently encoded in corresponding wordnets, and give some guidelines for their better coverage. Finally, we discuss the possibility to automatically generate new synsets and/or new relations on the basis of the most frequent and most regular derivational patterns.

Keywords: global wordnet, morpho-semantic relations, derivational relations

1 Introduction

The aims of this paper are to present the current stage of the encoding of morpho-semantic relations in Bulgarian and Serbian wordnets, briefly to sketch the derivational properties of Slavic languages based on the observations from Bulgarian and Serbian, to discuss the nature of morpho-semantic relations and its reflection to the wordnet structure and to analyze the positive and negative consequences of an automatic insertion of Slavic derivational relations into it.

The wordnet is a lexical-semantic network which nodes are synonymous sets (synsets) linked with the semantic or extralinguistic relations existing between them [3], [8]. The wordnet structure also includes semantic and morpho-semantic relations between literals (simple words or multiword expressions) constituting the different synsets. The representation of the wordnet is a graph. The cross-lingual nature of the global wordnet is provided by establishing the relation of equivalence between synsets that express the same meaning in different languages [15].

The global wordnet offers the extensive data for the successful implementation in different application areas such as cross-lingual information and knowledge management, cross-lingual content management and text data mining, cross-lingual information extraction and retrieval, multilingual summarization, machine translation, etc. Therefore the proper maintaining of the completeness and consistency of the global wordnet is an important prerequisite for any type of text processing to which it is intended.

The structure of the paper outlines the underlined goals. In the following section we present a short analysis of related work. In the third section, we briefly describe the properties of Slavic derivational morphology based on examples from Bulgarian and Serbian and their reflection into the wordnet structure. The forth section explains how the morpho-semantic relations are encoded in Bulgarian and Serbian wordnets respectively. We then discuss the manners to incorporate the (Slavic) derivational relations into the wordnet structure and some limitations of their automatic insertion. Finally, we raise some problematic questions connected with the presented study and propose future work to be done.

2. Related work

Wordnets have been developed for the most of the Slavic languages – Bulgarian, Serbian, Czech, Russian, Polish, Slovenian, and some initial work has been done for Croatian. Wordnets for three Slavic languages (Czech – started with the EuroWordNet (EWN), Bulgarian and Serbian) have been developed in the scope of the Balkanet project (BWN) [2], [11] and later on continue developing as nationally funded projects[1] or on the volunteer basis.

Originally, the Princeton wordnet (PWN) is designed as a collection of synsets that represent synonymous English lexemes which are connected to one another with a few basic semantic relations, such as hyponymy, meronymy, antonymy and entailment [3], [8]. This same structure has basically been mirrored in most of the wordnets developed on the basis of PWN. The structural difference of Slavic languages which show many similar features has induced the enrichment of wordnets with new information. Added information is mostly related to the inflectional and derivational richness of a language in question. For instance, information related to inflectional properties has been added to all lexemes in Bulgarian [4] and Serbian [2] wordnets, and for Serbian some rudimentary semantic relations that can be inferred from the derivational connectedness, for instance *derived-pos* (for possessive adjectives) and (for gender motion) *derived-gender* [2] has been added too. On the other hand, the recognized importance of PWN, and global wordnet in general, for various NLP applications has initiated the major additions and modifications of PWN itself.

The existence of derivational relations that exhibit a fairly regular behavior and that connect lexemes that belong to the same or to the different categories seemed to many as a good starting point for the substantial wordnet enrichment. We will present

[1] http://dcl.bas.bg/bulNet/general_en.html

the most interesting approaches. All these approaches rely on the fact that if there is a derivational relation between two lexemes belonging to different synsets then most probably there is a kind of semantic relation between the synsets to which the lexemes belong.

The automatic enrichment of wordnet on the basis of the derivational relations has been proposed and used for the Czech wordnet [10]. The basic and most productive derivational relations in Czech have been included in a Czech morphological analyzer and generator, and semantic labels were added to the derivational relations.

The sharing of semantic information across wordnets has been proposed by [1]. Namely, if wordnets for several languages are connected to each other (for instance via Interlingual index (ILI) [15], as has been done for wordnets developed in scope of EuroWordnet and Balkanet projects), then semantically related synsets in a source language for which the connection has been established on the basis of the derivational relatedness of some of the lexemes can be used to connect the synsets in a target language whose lexemes may not exhibit any derivational relation.

The method to improve the internal connectivity of PWN has been proposed in [9]. The existing synsets have been manually connected on the basis of the automatically produced list of pairs of lexemes that are (potentially) derivationally, and therefore also semantically, connected. In this paper we will try to show why we find the last approach the most appropriate for Bulgarian and Serbian.

3. Slavic derivation in wordnet structure

The derivation is highly expressive in all Slavic languages. Some of the most frequent and regular derivational mechanisms in Bulgarian and Serbian are given in the Table 1. The status of the derivational mechanisms listed is not the same. Some of them represent the more or less frequent models which are not applicable to every lemma that has certain syntactic or semantic property, while the other models can always be applied. For instance, the pattern **Verb → Noun representing a profession** is one of the numerous derivational pattern in Bulgarian (уча → учител) and Serbian (učiti → učitelj), while the pattern **Verb → Verbal noun** is a general rule that can be applied to all imperfective verbs in the two languages. Similarly, a possessive adjective exists for every animate noun [12]. We call this phenomenon a regular derivation since in some respect it enhances the notion of inflectional class.

Formally, regular derivation is performed by derivational operators that significantly influence the structuring of the lexicon of Slavic languages. The analysis of this phenomenon is given in [13], [14] on the examples of processing of possessive and relational adjectives, amplification and gender motion in various English-Serbian and Serbian-English dictionaries. Moreover, the derivational potential is, as a rule, connected to the specific sense of a lemma (see sections 5 and 7).

Table 1. Some of the derivational mechanisms in Bulgarian and Serbian

Relation	Bulgarian	Serbian	English

Aspect pairs	уча → научa	učiti → naučiti[2]	teach – learn
Verb → noun	уча → учител	učiti → učitelj	teach – teacher
Verb → noun	уча → ученик	učiti → učenik	learn – student
Verb → noun	уча → училище	učiti → učilište[3]	learn – school
Verb → noun		učiti → učionica	learn – classroom
Verb → noun	уча → учебник	učiti → udžbenik	learn – textbook
Verb → noun	уча → учен	učiti → učenjak	learn – scientist
Verbal noun	уча → учение	učiti → učenje	learn – studies
Verbal noun	уча → учене		learn – study
Collective noun	ученик → ученичество		student – schooldays
Verb → adjective	уча → учебен	učiti → učen	learn – educational
Verb → adjective	уча → учен	učiti → učevan	learn – educated
Relative adjective	учител → учителски	učitelj → učiteljski	of or related to teacher
Possessive adjective	учител → учителски	učitelj → učiteljev	male – female teacher
Gender pairs	учител → учителка	učitelj→ učiteljica	teacher – female teacher
Gender pairs		učiti → učenica	student -female student
Diminutive	ученик – учениче	učenik – učeničić	student – little student

4. Current state of morpho-semantic relations in Bulgarian and Serbian wordnets

Eight semantic relations between synsets are represented (in a correspondence with the Princeton wordnet) in Bulgarian [4], [5] and Serbian wordnets [2].These relations are: hypernymy, meronymy (three subtypes are registered among others recognized), subevent, caused, be in state, verb group, similar to and also see (also see in PWN actually encodes two different relations: between verbs and between adjectives, the former one being a kind of morpho-semantic relation between literals roughly corresponding to Slavic verb aspect while the second one is a semantic relation of similarity between synsets). Three extralinguistic relations between synsets are encoded as well: usage domain, category domain and region domain. The wordnet structure includes also semantic and morpho-semantic (derivational) relations among literals belonging to the same or to the different synsets. Semantic relations between literals are: synonymy and antonymy (in Bulgarian and Serbian wordnets antonymy links synsets); derivational are: derived, participle, derivative in Bulgarian, and derived-pos, derived-gm, and derived-vn in Serbian.

4.1 Encoded morpho-semantic relations

The morpho-semantic relations in Bulgarian and Serbian wordnets link synsets although they derivationally apply to the literals only (single word and multi-word lemmas). On the other hand, morpho-semantic relations express different kinds of semantic relations which hold between synsets. Neither the derivational links between the exact literals, nor labels [10] for the respective semantics relations operating between synsets are encoded so far in Bulgarian and Serbian wordnets. The subsumed

[2]There is actually a whole list of perfective verbs that correspond to the imperfective verb учити: izučiti, naučiti, obučiti, preučiti (se), podučiti, poučiti, priučiti, proučiti.

[3] Today obsolete.

morpho-semantic relations are briefly presented below (some statistical data are shown in Table 2):

Derivative is an asymmetric inverse intransitive relation between derivationally and semantically related noun and verb. For example the Bulgarian literal **водя** from the synset {насочвам:1, насоча:1, **водя:4**, напътвам:1, напътя:1, направлявам:1} (the corresponding English synset is {steer:1, maneuver:1, maneuver:2, manoeuvre:2, direct:11, point:4, head:5, guide:1, channelize:1, channelise:1} with a definition '*direct the course; determine the direction of traveling*') is in *derivative* relation with the noun **водач** from the synset {**водач:3}** (the corresponding English synset is {guide:2} with a meaning '*someone who shows the way by leading or advising*').

Derived is an asymmetric inverse intransitive relation between derivationally and semantically related adjective and noun. For example the literal **меден** from the Bulgarian synset {меден:1} (the English equivalence {cupric:1, cuprous:1} with a definition '*of or containing divalent copper*') is in a *derived* relation with the literal **мед** from the synset {**мед**:2, Cu:1} (in English → {copper:1, Cu:1, atomic number 29:1}). A productive derivational process rely Slavic nouns with respective relative adjectives with general meaning '*of or related to the noun*'. For example, the Bulgarian relative adjective {*стоманен*:1} defined as '*of or related to steel*' has the Serbian equivalent {*čelični*:1} with exactly the same definition. Actually in English this relation is expressed by the respective nouns used with an adjectival function (rarely at the derivational level, consider wooden↔wood, golden↔gold), thus the concepts exist in English as well and the mirror nodes should be envisaged.

Participle is an asymmetric inverse intransitive relation between derivationally and semantically related adjective denoting result of an action or process and the verb denoting the respective action or process. Consider **играя** from {**играя**:7} (the English equivalent {play:1} with a definition '*participate in games or sport*') which is in a *Participle* relation with the literal **игран** from {**игран**:1} denoting '*(of games) engaged of*' for the English counterpart {played:1}. All Bulgarian verbs produce participles (the number of participles varies from one to four depending on the properties of the source verb) which are considered as verb forms constituting complex tenses or passive voice. On the other hand, a big part of the Bulgarian participles acts as adjectives with separate meaning. The similar relations between a verb and its participles hold for Serbian.

It can be seen that the actual derivational relations are established between particular literals although the synsets are formally linked (the actual semantic relation between synsets which marker is the derivation itself is not labeled). The English *derivative, derived,* and *participle* relations are automatically transferred to Bulgarian wordnet. As they are language specific and obviously there is no one to one mapping between English and Bulgarian the expanded links are manually validated. A specification whether a given morpho-semantic relation exists in English only is declared in a synset note (SNote).

The relation *eng_derivative* has been also automatically transferred to Serbian although the corresponding derivational relation may hold in Serbian as well but need not (see the Serbian example in section 5). The new relations *derived-pos*, *derived-vn*, and *derived-gender* have been introduced in Serbian wordnet to relate possessive and relative adjectives, verbal nouns and female (or male) doublets, assigned mainly to the Balkan specific or Serbian specific synsets.

Table 2. Statistical data for the encoded morpho-semantic relations in Bulgarian and Serbian wordnets.

Number of	BG WN	SR WN	PWN 2.0
Synsets	29,136	13,612	115,424
Literals	56,223	23,139	203,147
Relations	53,144	18,210[4]	204,948
Derived	1,696	314	1,296
Derivative	8,920	83[5]	36,630
Participle	212	0	401

4.2. Not-encoded morpho-semantic relations

The general observations are that not all existing *derivative, derived,* and especially *participle* links are marked in Bulgarian and Serbian wordnets. The main reason originates in the language specific character of the word-building in view of the fact that an exact correspondence with the PWN has been mostly followed in the expand wordnet model. As a result a lot of language-specific derivational relations (that can be described in terms of *derivative, derived,* and *participle* relations) remain unexpressed in Bulgarian and Serbian wordnets. For example the literals from the Bulgarian synset {метален:1, металически:1} corresponding to the English synset {metallic:1, metal:1} with a definition: '*containing or made of or resembling or characteristic of a metal*' are derived from the literal **метал** from the synset {**метал**:1, метален елемент} equal to the English synset {metallic element:1, metal:1} with a definition '*any of several chemical elements that are usually shiny solids that conduct heat or electricity and can be formed into sheets etc*'. Nevertheless the corresponding *derived* relation is not linked in the Bulgarian wordnet. Consider the following more complicated example. The literal **пекар** from the Bulgarian synset {**пекар**:1, **хлебар**:1, **фурнаджия**:1} (English equivalent {baker::2, bread maker:1} with a definition '*someone who bakes bread or cake*') is in a *derivative* relation with the literal **пека** from the synset {**пека**:1, опичам:1, опека:1, изпичам:1, изпека:1} (in English {bake:1} with a definition '*cook and make edible by putting in a hot oven*'). Moreover the second target literal **хлебар** is in a derivational relation with the source literal **хляб** from the synset {**хляб**:1} (in English {bread:1, breadstuff:1, staff of life:1} with a definition '*food made from dough of flour or meal and usually raised with yeast or baking powder and then baked*'), while the third one **фурнаджия** is in a derivational relation with the source literal **фурна** from {пекарница:1, **фурна**:2} (in PWN {bakery:1, bakeshop:1, bakehouse:1} with a definition '*a workplace where baked goods (breads and cakes and pastries) are produced or sold*'). None of the three existing derivational relations is encoded in the Bulgarian wordnet so far.

In Serbian, for instance, the adjective synset {**zamisliv**:1} (English equivalent is {conceivable:2, imaginable:1, possible:3} with a definition *'possible to conceive or imagine*') is not linked with the verbal synset {**zamisliti**:2y, koncipirati:1b} (in

[4] Without extralinguistic relations: category and region, and relation eng_derived.
[5] Includes relations: *derived-pos*, *derived-vn*, and *derived-gender*.

English {imagine:1, conceive of:1, ideate:1, envisage:1} with a definition '*form a mental image of something that is not present or that is not the case*'), although relation *derived*, or some more specific, would be appropriate.

4.3. Language-specific morpho-semantic relations

There are systematic morpho-semantic differences concerning derivational mechanisms between English and Slavic languages [7]. Some of the most productive derivational relations in Slavic languages are briefly presented here: namely verbal aspect pairs, gender pairs, and diminutives.

4.3.1. Aspect pairs

The verb aspect is a category that occurs in all Slavic languages, its nature is very sophisticated. Generally speaking, the verb aspect in Slavic languages can be descried as a relation between the action and its bound (limit) regardless of the person, speaker and speech act. The perfect aspect verbs express integrity and completeness, while the imperfect aspect verbs – lack of integrity or a process (duration, recurrence). Each Slavic verb is either perfective or imperfective; there are a number of verbs that are bi-aspectual and act as both imperfective and perfective. Most verbs form strict pairs where perfective and imperfective members form a derivational relation between two lexemes expressing generally the same meaning. The Bulgarian verbs are classified as: imperfective (perfective correspondent exists), perfective (imperfective correspondent exists), bi-aspectual, imperfective tantum (perfective correspondent does not exist), perfective tantum (imperfective correspondent does not exist). In Bulgarian wordnet the aspect pairs are introduced in one and the same synset with an LNote (literal note) describing the respective aspect. For example {съчинявам:2 LNOTE: imperfective, съчиня:2 LNOTE: perfective, пиша:4 LNOTE: imperfective, написвам:2 LNOTE: imperfective, напиша:2 LNOTE: perfective} (an equivalent of the English synset {write:1, compose:3, pen:1, indite:1} with a definition '*produce a literary work*'). Similarly, in Serbian wordnet the aspect pairs are introduced in a same synset. For instance in a synset {**zamišljati**:2x, **zamisliti**:2x, dočaravati:2x, dočarati:2x, predočavati:1, predočiti:1} (in English {visualize:1, visualise:3, envision:1, project:9, fancy:1, see:4, figure:3, picture:1, image:1} with a definition '*imagine; conceive of; see in one's mind*'), LNOTE element corresponding to each literal describes inflectional and derivational properties of each verb, e.g. LNOTE content for the imperfective verb **zamišljati** is V1+Imperf+Tr+Iref+Ref, while LNOTE content the perfective correspondent **zamisliti** is V162+Perf+Tr+Iref+Ref [6]. In most cases, however, perfective verbs derived from the imperfective by prefixation express different meaning and are not in the same synset, for example the perfective verb **uraditi** '*do, perform*' and its imperfective correspondent **raditi** are not in the same synset.

4.3.2. Gender pairs

The gender pairing is systematic phenomenon in Slavic languages that display binary morpho-semantic opposition: male → female, and as a general rule there is no

corresponding concept lexicalized in English. The derivation is applied mainly to nouns expressing professional occupations, but also to female (or male) correspondents of nouns denoting representatives of animal species. For example, Bulgarian synset {преподавател:2, учител:1, инструктор:1} and Serbian synset {predavač:1} that correspond to the English {teacher:1, instructor:1} with a definition: '*a person whose occupation is teaching*' have their female gender counterparts {преподавателка, учителка, инструктурка} and {predavačica} with a feasible definition '*a female person whose occupation is teaching*'.

There are some exceptions where like in English one and the same word is used both for masculine and feminine in Bulgarian and Serbian, for example {президент:1} which corresponds to the English synset {president:3} with a definition: '*the chief executive of a republic*', and as a tendency the masculine noun can be used referring to females. Following the PWN practice the female counterparts are encoded in Bulgarian and Serbian wordnets as hyponyms of the corresponding synset with the male counterpart. For example {актриса:1} (English equivalent {actress:1} with a definition '*a female actor*') is a hyponym of {актьор:1, артист:} (corresponding to the English synset {actor:1, histrion:1, player:3, thespian:1, role player:2} expressing the meaning '*a theatrical performer*'). It might be foreseen of introducing a new relation describing the female – male opposition of nouns in Slavic languages as has already been done for Serbian.

4.3.3. Diminutives

Diminutives are standard derivational class for expressing concepts that relate to small things. The diminutives display a sort of morpho-semantic opposition: big → small, however sometimes they may express an emotional attitude too. Thus the following cases can be found with diminutives: standard relation big → small thing, consider {стол:1} corresponding to English {chair:1} with a meaning '*a seat for one person, with a support for the back*' and {столче:1} with an feasible meaning '*a little seat for one person, with a support for the back*'; small thing to which an emotional attitude is expressed. Also, Serbian synset {lutka:1} that corresponds to the English {doll:1, dolly:3} with a meaning '*with a replica of a person, used as a toy*' is related to {lutkica} which has both diminutive and hypocoristic meaning. There might be some occasional cases when this kind of concept is lexicalized in English, {foal:1} with a definition: '*a young horse*', {filly:1} with a definition: '*a young female horse under the age of four*', but in general these concepts are expressed in English by phrases.

For the moment the diminutives are included in Bulgarian and Serbian wordnets only in the rare case when the English equivalent is lexicalized. On the other hand, almost from every concrete noun a diminutive (in some cases more than one lexeme) can be derived. Consequently a place for the diminutives in the wordnet structure has to be provided.

5. The nature of morpho-semantic (derivational) relations

One of the most important features of the morpho-semantic relations is that being derivational relations between literals (i.e. assistant is a person that assists; participant is the person that participates etc.) they express also regular semantic oppositions holding between synsets [9]. The derivational relation linking **assist** and **assistant** from the respective synsets {help:1, **assist**:1, aid:1} '*give help or assistance; be of service*' and {**assistant**:1, helper:1, help:4, supporter:3} '*a person who contributes to the fulfillment of a need or furtherance of an effort or purpose*' implies a kind of semantic relation over synsets formulated in [10] as an agentive relation existing between an action and its agent.

Given morpho-semantic relation may be realized by different derivation mechanisms. Consider the literals from the Bulgarian synset {певец:2, вокалист:1} (in English {singer:1, vocalist:1, vocalizer:2, vocaliser:2} with a definition '*a person who sings*'), the former one **певец** is derived with the suffix –**ец** from the literal **пея** constituting the synset {пея:1} (the English equivalent {sing:2} with a definition '*produce tones with the voice*'}, while the second one **вокалист** is derived with the suffix –**ист** from the literal **вокализирам** belonging to the synset {вокализирам:1} (in English {vocalize:2, vocalise:1" with a definition '*sing with one vowel*').

On the other hand, different derivational mechanisms might correspond to different semantic relations. For example in Bulgarian, as well as in English the verb **чета** from the synset {**чета**:3; прочитам:2; прочета:2} corresponding to the English synset {**read**:1} with a definition '*interpret something that is written or printed*' has the following derivates among others:

– the noun **четене** from the synset {**четене**:1} ↔ {**reading**:1}, with a definition: '*the cognitive process of understanding a written linguistic message*'. The derivation transforms the verb into a verbal noun. The respective relation between synsets is formulated as an action relation in [10].

– the noun **читател** from the synset {**читател**:1} ↔ {**reader**:1}, with a definition: '*a person who enjoys reading*'. The derivational relation links the source verb with a noun build by an affixation. The respective relation between synsets expresses a property over the underlying action.

In some cases when the source literal has more than one meaning the exact correspondences with the derivates can be traced. Consider the verb **чета** from the synset {**чета**:1, прочитам:1, прочета:1} equivalent with the English synset {read:3} with a definition '*look at, interpret, and say out loud something that is written or printed*'. Its verbal noun derivative **четене** from the synset {**четене**:1; поетическо четене:1; рецитал:1} (in English {recitation:2, recital:3, reading:7}) expresses a meaning which is related with the meaning of the source '*a public instance of reciting or repeating (from memory) something prepared in advance*'. As the source derivates counterpart in two different synsets (equivalent to {read:1} and {read:3}), this presupposes the corresponding difference in the meanings of the resulting derivatives. Thus the same derivational mechanism might indicate for different semantic oppositions if it targets graphically equivalent literals expressing different meaning (the observed difference in the semantic oppositions remains undistinguished). It is natural that the synsets {read:1} and {read:3} are related with a *verb group* relation.

The semantic part of the morpho-semantic relations is not language specific, language specific are the derivational mechanisms of lexicalization. There are several English derivatives of the literal **paint** from {paint:3} with a definition '*make a painting of*':

En 1.{**paint**:1} – '*a substance used as a coating to protect or decorate a surface (especially a mixture of pigment suspended in a liquid); dries to form a hard coating*'

En 2. {**painter**:1} – '*an artist who paints*'

En 3. {**painting**:1, picture:2} – '*graphic art consisting of an artistic composition made by applying paints to a surface*'

En 4.{**painting**:2} – '*creating a picture with paints*'

Neither of the corresponding Bulgarian equivalents:

Bg 1.{боя:2}

Bg 2.{живописец:1, художник:1}

Bg 3.{картина:3}

Bg 4,{живопис:1}

are derivatives of the Bulgarian synset equivalent to {paint:3} – {рисувам:2; нарисувам:2}. Nevertheless the same semantic oppositions exist in Bulgarian although they are not marked with any semantic or morpho-semantic relations.

In Serbian some of the related synsets to {naslikati:1 LNOTE: V101+Perf+Tr+Iref} (equal to {paint:3}) include derivatives, while the other do not (e.g. {boja:2x, farba:1x}). The *derivative* relation is transferred from English to Serbian wordnet, but the name of the relation has not been changed in order to indicate that the origin of the relation is English, and that it may hold for Serbian but need not, as shown by the same example.

Sr 1. {boja:2x, farba:1x}

Sr 2. {slikar:1}

Sr 3. {slika:1}

Sr 4. {slikarstvo:1}

This means that the derivational relations in a particular language might be successfully used not only for the detecting of a given semantic opposition. Moreover they can be exploited for the identification of the corresponding semantic relations in other languages where lexicalization is expressed by different mechanisms. Thus we have to make clear distinction between the derivation as a literal relation (asymmetric, inverse, and intransitive) and the semantic oppositions between synsets for which the derivation itself might be a formal pointer.

6. Approaches to cover Slavic specific derivations in wordnet

There are several possible approaches for covering different lexicalizations resulting from derivation in different languages [7], [11]:

– to treat them as denoting specific concepts and to define appropriate synsets (gender pairs in Bulgarian and Serbian; relative adjectives in Bulgarian and Serbian);

– to include them in the synset with the word they were derived from (verb aspect in Bulgarian and in most of the cases in Serbian);

– to omit their explicit mentioning (diminutives in Bulgarian);

– to provide source literals with flexion-derivation description encompass these phenomena as well.

Treating morpho-semantic relations such as verb aspect, relative adjectives, gender pairs and diminutives among others in Slavic languages as relations that involve language specific concepts requires an ILI addition for the languages where the concepts are presented (respectively lexical gaps in the rest). This solution takes grounds from the following observations:

– Verb aspect pairs, relative adjectives, feminine gender pairs and diminutives denote an unique concept;

– Verb aspect pairs, relative adjectives, feminine gender pairs and diminutives are lexicalized with a separate word in Bulgarian, Serbian, Czech and other Slavonic languages;

– Relative adjectives, feminine gender pairs and diminutives in most of the cases belong to different category or different inflectional class comparing to the word from which they are derived (there are some exceptions in the difference of the category, like diminutives that are derived from neuter nouns in Bulgarian).

Although the new wordnets do not compare yet with PWN's coverage, the former are continuously extended and improved so that a balanced global multilingual wordnet is foreseen. For that reason the task of proper encoding of different levels of lexicalization in different languages is becoming more and more important in the view of the various Natural Language Processing tasks. The Slavic languages possess rich derivational morphology which has to be involved into the strict one-to-one mapping with the ILI.

7. Automatic building of derivational relations in Bulgarian and Serbian

The derivational relations for literals that already exist in wordnet can be interpreted in terms of derivational morphology, e.g., the noun *teacher* is derived from the verb *teach* and so on. Wordnet already contains a lot of words that are produced by the derivational morphology rules: verbal nouns are linked with verbs, etc. In order to make explicit the morpho-semantic relations that exist already it would be necessary to include more links. On the other hand, a special attention has to be paid on the language specific derivational relations (some of them valid for big language families as Slavic languages). Several problems can be formulated following the observations and analyses presented in this study:

It is necessary to distinguish the pure derivation form the semantic relations which meaning is presupposed by the derivation itself. Concerning Bulgarian and Serbian wordnets this will be reflected particularly in the proper encoding of the derivational links between exact literals as it has been done in PWN; in the identification of derivational relations between literals already encoded in wordnets (comparing with PWN or exploiting language-specific derivational models), and in the introducing of language specific derivations in their appropriate place in the wordnet structure providing the exact correspondence with other languages.

In more general plan a theoretical investigation is needed to describe the nature of the semantic relations to which derivations are formal pointers. Ones a consistent classification is provided the respective semantic relations might be identified in the wordnets on the basis of the derivational ones in a particular language

Several tasks may be done semi-automatically: to link literals instead of synsets with derivational relations; and to identify synsets where the potentially derivationally related literals appear. Bellow we provide some observations why the complete automation is not appropriate; although the derivational regularities are in most of the cases well established.

Although derivation is in many cases regular in the sense that it yields predictable results, it cannot be freely used for generation since it can lead to over-generation; namely, one could generate something which exists in a language system but does not exist in language usage. For instance, in Bulgarian and Serbian an abstract noun can be regularly derived (with a suffix *–ocm*; *–ost*) from a descriptive adjective *X* meaning '*the quality of something that has the characteristic X*', and a prefix (*–не*, *–не*; *–без*, *–bez*, etc.) can be used to produce both the adjective and a noun with the opposite meaning. One such example in Serbian is *osećajan* '*be able to respond to affective changes*' → *osećajnost* '*the ability to respond to affective changes*' → *bezosećajan* '*not being able to respond to affective changes*' → *bezosećajnost* '*the inability to respond to affective changes*'. However, if the same pattern is applied to the adjective *sličan 'marked by correspondence or resemblance*' → *sličnost* '*the quality of being similar*' → ?*nesličan* '*not similar*' → ?*nesličnost* '*the quality of being dissimilar*', two last lexemes in a sequence though easily understood are not lexicalized.

In a context of a wordnet production it is not sufficient to produce new synsets, for instance by applying the regular derivational mechanisms. It is equally important to place the generated synsets in the already existent network consisting of various relations. For instance, in Serbian the nouns **sposobnost**, **vidljivost** and **popustljivost** are regularly generated from the adjectives **sposoban** '*having the necessary means or skill to do something*', **vidljiv** '*having the characteristics that make it visible'* and **popustljiv** *'easily managed or controlled*'. However, the produced nouns have three different hypernyms: *osnovna karakteristika* {quality:1}, *svojstvo* {property:3}, and *osobina* {trait:1}. The correct placement of newly generated synsets in an existent network is not straightforward.

It has been noted (in section 5) that many senses of some words are distinguished by their different derivational capabilities. For instance, Serbian verb **polaziti** has five different meanings according to the Serbian explanatory dictionary, and one submeaning of the second presented meaning is '*to go somewhere regularly and often to perform some duty*'. That meaning is the only one from which the noun **polaznik** '*someone who attends a school or a course*' can be derived by the agentive relation (realized by a suffix *–ik*).

It has already been stated in [8] that even derivation that seems very predictable can show very unpredictable behavior. Some derivational mechanisms in Bulgarian and Serbian are very predictable, like production of possessive adjectives that are produced from (mostly) animate nouns. As a consequence possessive adjectives are not listed in traditional Serbian dictionaries. The production of verbal nouns from imperfective verbs is also regular and produces a predictive meaning, the act of doing something. The verbal nouns are however, listed as a separate entries in Bulgarian and

Serbian dictionaries. Besides the predicted meaning they often acquire the additional meaning. For instance, the verbal nouns **ученuе** in Bulgarian, **učenje** in Serbian and *pečenje* in Serbian are derived from imperfective verbs **уча**, **učiti** 'to study' and **peći** 'to roast'. Besides the predicated meanings '*the act of studying*' and '*the act of roasting*' they have acquired in Serbian the additional meaning, 'doctrine' and 'roast meat', respectively. In the case of other derivational mechanism it can be more difficult to establish the meaning of the derived word. For instance, adjectives **pričljiv** and *čitljiv* in Serbian are derived respectively from the verbs **pričati** 'to talk' and **čitati** 'to read' using the same suffix *–iv*. Both verbs are imperfective and can be used both as transitive and intransitive: *Marko priča priču* 'Marko tells the story', *Marko puno priča* 'Marko speaks a lot', *Puno ljudi čita knjigu* 'A lot of people read the book', *Marko puno čita* 'Marko reads a lot'. The meaning of the adjective **pričljiv** is derived from the itransitive usage of a verb (namely, *Marko puno priča* implies *Marko je pričljiv* 'Marko is talkative'), while the adjective **čitljiv** is derived from the transitive usage (here *Puno ljudi čita knjigu* implies *Knjiga je čitljiva* 'The book is easy to read').

The complexity of the issue of automation is best illustrated by the derivation of gender pairs in Serbian, since they exhibit all the previously mentioned problems. If we consider the derivation of female counterparts for the nouns of professions we encounter the following situations:

- The female counterpart morphologically does not exist: for instance, **sudija** *'judge'* is therefore used for both men and women;
- The female counterpart morphologically exists but is never used, **vojnik** *'solider'* vs. ***vojnica** and **žena vojnik** '*(woman) soldier*';
- The female counterpart exists and is exclusively used for women performing that profession or function: **kelner** *'waiter'* and **kelnerica** *'waitress'*;
- The female counterpart exists but the male noun is also sometimes used for women: **profesor** *'(man or woman) professor'* and **profesorka** '*(woman) professor*';
- The female counterpart exists but it does not mean quite the same as a noun it was derived from: **sekretar** '*secretary*' is treated as someone performing an highly responsible function, as opposed to **sekretarica** '*(woman) secretary*' who is performing the low-level tasks in an organization;
- The female counterpart exists but it has acquired a different meaning, so it is not used to denote a woman performing certain function: **saobraćajac** '*traffic cop*' vs. **saobraćajka** '*car accident*'.

8. Conclusions and future work

We have briefly presented the current stage of the encoding of morpho-semantic relations in Bulgarian and Serbian wordnets. Grounding on the derivational properties of Slavic languages we provided some observations over the sophisticated nature of morpho-semantic relations and presented some examples proving the negative consequences from a purely automatic insertion of Slavic derivational relations into the wordnet structure. We believed we added additional evidences supporting the approach presented in [9] namely the utilization of a semi-automatic identification or

insertion of morpho-semantic relations. Such an approach would significantly facilitate the wordnet development although a manual connection on the basis of the automatically produced lists of suggested pairs has to be provided.

Further development of both Bulgarian and Serbian wordnets is narrowly connected with an investigation towards the theoretical grounds of the nature of morpho-semantic relations. At first stage the encoding of derivational relations between exact literals instead of synsets is foreseen. Another important task is the introducing of Slavic language specific derivations in a uniform way providing at the same time ILI correspondences. The accomplishment of these tasks will also reflect in the successful implementations of approaches based on cross-lingual information extraction, retrieval, and data mining, multilingual summarization, machine translation, etc.

References

1. Bilgin, O., Cetinouğlu, O. and Oflazer, K. Morphosemantic Relations In and Across Wordnets – A Study Based on Turkish. In: Proceedings of the Global Wordnet Conference (Sojka P, et al), Brno 2004, pp. 60 –66. (2004)
2. Christodulakis, D. (ed.) Design and Development of a Multilingual Balkan Wordnet (BalkaNet IST-2000-29388) – Final Report. (2004)
3. Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press. (1998)
4. Koeva, S., 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. (2004)
5. Koeva, S., Bulgarian WordNet – development and perspectives, in: International Conference Cognitive Modeling in Linguistics, 4-11 September 2005, Varna. (2005)
6. Krstev, C., Vitas, D., Stanković, R., Obradović, I. and Pavlović-Lažetić, G. Combining Heterogeneous Lexical Resources. In: Proceedings of the Fourth Interantional Conference on Language Resources and Evaluation, Lisabon, May 2004, vol. 4, pp. 1103-1106. (2004)
7 Krstev, C., S. Koeva, D. Vitas, Towards the Global Wordnet, in: First International Conference of Digital Humanities Organizations (ADHO) Digital Humanities 2006, Paris-Sorbonne, 5-9 July 2006, pp. 114-117. (2006)
8. Miller G. R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller. Five Papers on WordNet. Special Issue in International Journal of Lexicography, vol.3, no.4. (1990)
9. Miller, G. A. and Fellbaum, C. Morphosemantic links in Wordnet. Traitement automatique des langues, 44.2:69-80. (2003)
10. Pala, K. and Hlavačková, D. Derivational Relations in Czech Wordnet. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, ACL, Prague, 75-81. (2007)
11. 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. (2002)
12. Vitas, D., Krstev, C. Regular derivation and synonymy in an e-dictionary of Serbian, Archives of Control Sciences, Polish Academy of Sciences, Volume 51(LI), No. 3, pp. 469-480. (2005)
13. Vitas, D. Morphologie dérivationnelle et mots simples: Le cas du serbo-croate, Lingvisticae Investigationes Supplementa 24 (Lexique, Syntaxe et Lexique-Grammaire / Syntax, Lexis &

Lexicon-Grammar - Papers in honour of Maurice Gross), John Benjamin Publ. Comp., pp. 629-640, (2004)
14. Vitas, D., Krstev, C. Restructuring Lemma in a Dictionary of Serbian, in Erjavec, T., Zganec Gros, J. (eds.) Informacijska druzba IS 2004" Jezikovne tehnologije Ljubljana, Slovenija, eds., Institut "Jozef Stefan", Ljubljana, 2004
16. Vossen P. (ed.) EuroWordNet: a multilingual database with lexical semantic networks for European Languages. Kluwer Academic Publishers, Dordrecht. (1999)