

# Derivational and Morphosemantic Relations in Bulgarian Wordnet

Svetla Koeva

Institute for Bulgarian, Bulgarian Academy of Sciences,  
Sofia, Bulgaria

## Abstract

In this paper we deal with the problem of representing the derivational and morphosemantic relations in wordnets, especially problems that arise in the process of developing, on the basis of the Princeton WordNet, wordnets for languages significantly different from English. In particular, we present how derivational relations are presently encoded in Bulgarian wordnet and give some directions for their better coverage. We also discuss the nature of derivational and morphosemantic relations and its reflection in the Bulgarian wordnet, as well as the proper encoding of different levels of lexicalization in different languages.

**Keywords:** wordnet relations, derivational and morphosemantic relations

## 1 Introduction

The Bulgarian wordnet (BulNet) has been in development for the past six years within the framework of the European project *BalkaNet* (Tufiş et al., 2004) and the project *BulNet – Lexical-Semantic Network of Bulgarian* (Koeva et al., 2004). At the beginning of 2008 BulNet includes 29,275 synsets (that is approximately one-fourth of the English wordnet). The number of unique literals is 44,238, which represents about two-third of a standard spelling dictionary.

Generally speaking, the derivational relations are language specific – either resulting in language-specific concepts (such as Slavic diminutives, augmentatives, etc.) or manifesting language-specific derivational properties (such as Slavic verbal nouns, participles, etc.). The number of derivational relations included in BulNet so far is not large enough, although Bulgarian is a language with rich derivational morphology. On the other hand, the importance of derivational relations in NLP is undoubted. Thus the proper explication of derivational relations will enrich density and connectivity in Bulgarian wordnet, providing at the same time unique language resource where derivational links are semantically related.

Some derivational relations express different kinds of morphosemantic relations, which hold between synsets. Recently, “morphosemantic links” that connect words derived by means of a morphological affix have been added to WordNet (Miller and Fellbaum, 2003). Moreover, it was suggested that the meanings of affixes can be classified into a relatively small number of semantic categories, labeled as agent, instrument, etc. (Fellbaum et al., 2007; Pala and Hlavachkova, 2007).

The goals of this paper are: to outline the description of natural language semantic relations in terms of binary relations over a set, with respect to the Bulgarian wordnet; to present the current state of the encoding of derivational relations in BulNet and to propose guidelines for their better coverage; to discuss the nature of derivational and morphosemantic relations and its reflection in the Bulgarian wordnet; as well as to pay attention to the proper encoding of different levels of lexicalization in different languages. The structure of the paper outlines these goals. In this study, morphosemantic and derivational relations are not used as synonymous terms, since the first one expresses a semantic relation between synsets, which might be indicated by a given derivational relation between literals, in its turn.

## 2 Language internal relations

The semantic relations in wordnet reflect factual real-world relations (among sets of objects or abstractions). The description of wordnet relations in terms of binary relations over a set is important for wordnet validation in order to preserve the consistency of the structure and the uniformity of the representation (equivalent relations are linked automatically, transitive relations are checked for transitive loops, etc.). For example, once the hypernymy has been defined as an inverse (to hyponymy) transitive semantic relation between synsets (nouns or verbs) that expresses inclusion of classes, no occurrences of multiple hypernymy should be allowed. There are still some cases of multiple hypernymy in the Princeton WordNet (PNW), hence in BulNet, as well, because at this stage hypernymy combines various types of inclusions, e.g.  $\{oxygen:1, O:2, atomic\ number\ 8:1\}$  has two hypernyms:  $\{chemical\ element:1, element:2\}$  and  $\{gas:2\}$  although a separate type of the hypernymy relation between noun synsets (known as an instance hypernymy) is defined to cover different instances of one and the same concept. All relations between objects in Bulgarian wordnet – literals and synsets – are described in terms of well-known equivalent, inverse and transitive binary relations; for the entire description of wordnet relations properties, the Euclidean relation is also used, as well as “multiple” and “domain” relations, which are formulated as follows.

A relation  $R$  within a given set  $X$  is Euclidean, if, for each  $x$ ,  $y$ , and  $z$  from  $X$ ,  
whenever both  $xRy$  and  $xRz$  are valid, so is  $yRz$ .

A relation  $R$  within a given set  $X$  is multiple, iff, for each  $x$ , there are at least  
two individuals  $y$  and  $z$ , for which  $xRy$  and  $xRz$  are valid.

A domain relation  $E$  within a given set  $X$  is present if, when the relation  $xEy$  is  
true for each  $x$  from  $X$ , then for each hyponym  $z$  of  $x$  it is true that  $zEy$ .

The formal representation of the relations supports the formulation of the wordnet descriptive logic, which, on the other hand, is very powerful in the applications for validation of wordnet completeness and consistency (Koeva et al., 2004). The search engine of Hydra, a tool for BulNet editing, viewing and validating, is based on this wordnet modal language. Provided that a given wordnet property is definable as a formula in the modal language, the tool determines all the objects in

the wordnet structure validating the formula, and hence the property, covering an automatic consistency validation.

## 2.1 Semantic relations between synsets

The semantic relations represented so far in BulNet (in correspondence to PWN) are grouped according to their properties as follows:

**Inverse and transitive:** hypernymy

**Inverse, transitive and multiple:** hyponymy, meronymy (three subtypes are used among others recognized), holonymy, has subevent, is subevent of

**Inverse and multiple:** has attribute

**Inverse:** is attribute of, causes, is caused by

**Symmetric, Euclidian and multiple:** similar to, verb group, also see

**Also see** in PWN actually encodes two different relations: between verbs and between adjectives, the former being a kind of derivational relation between literals roughly corresponding to Bulgarian verb aspect, the latter being a semantic relation of similarity between synsets.

## 2.2 Literal relations

The wordnet structure includes also semantic and derivational relations among literals belonging to the same or to different synsets. The semantic relations between literals are: synonymy (a reflexive, symmetric and transitive semantic relation of equivalence between literals within a synset) and antonymy (a symmetric semantic relation of opposition between literals from different synsets that belong to one and the same part of speech). In the Bulgarian wordnet antonymy links synsets and is called near-antonymy following the EuroWordNet specifications. Encoded derivational relations are: has a derived (derived), has a participle (participle), has a derivative (derivative). Recently, derivational relations have been distinguished from the so-called morphosemantic relations, such as agent, instrument, etc., which are semantic relations themselves and which might be indicated by derivational relations.

## 3 Derivational relations

The derivational relations in the Bulgarian wordnet follow the BalkaNet framework link synsets, although they derivationally apply to the literals only to the extent this has been done in PWN. The English derivative, derived, and participle relations are automatically transferred to the Bulgarian wordnet. As they are language-specific and, obviously, there is no one-to-one mapping between English and Bulgarian, the expanded links have been validated manually. A specification whether a given derivational relation exists in English only is declared in a literal note (LNote).

### 3.1 Derivational relations encoded so far in the Bulgarian wordnet

**Has a derived (derived)** is an inverse (has a derived could be multiple, as well) derivational relation between a noun and an adjective formed from it, e.g.: the literals from the Bulgarian synset {психологичен:1, психологически:3} – {psychological:2}: ‘of or relating to or determined by psychology’ are in a derived relation to the literal from the synset {психология:1} – {psychology:1, psychological science:1}. The derivation linking nouns with respective relative adjectives with the general meaning ‘of or related to the noun’ is very productive in Bulgarian.

**Has a participle (participle)** is an inverse (has a participle could be multiple, as well) derivational relation between a verb and a participle, formed from it, e.g.: each literal from the synset {счукан:1, смлян:1, стрит:1, смачкан:1} – {crunched:1}: ‘reduced to small pieces’ is a passive participle of the respective verbs from the synset {счукам:1, смилам:1, струвам:1, смачкам:5} – {grind:3, mash:3, crunch:4, bray:2, comminute:1}: ‘reduce to small pieces or particles by pounding or abrading’. All Bulgarian verbs produce participles (the number of participles varies from one to four depending on the properties of the source verb) which act either as verb forms, in the formation of complex tenses, passive voice and conditional mood, or as adjectives.

**Has a derivative (derivative)** is an inverse derivational relation (has a derivative could be multiple, as well) between a verb and a noun formed from it, e.g. the Bulgarian literal вода from the synset {насочвам:1, вода:4, направлявам:1} – {steer:1, maneuver:1, maneuver:2, manoeuvre:2, direct:11, point:4, head:5, guide:1, channelize:1, channelise:1}: ‘direct the course; determine the direction of traveling’ is in derivative relation with the noun водач from the synset {водач:3} – {guide:2}: ‘someone who shows the way by leading or advising’.

### 3.2 Not-encoded derivational relations

The general observations are that not all existing derivative, derived, and especially participle links are marked in BulNet. The main reason lies in the language-specific character of word-building in view of the fact that an exact correspondence to PWN has been, at most part, 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. For example, the literals from the Bulgarian synset {метален:1, металически:1} – {metallic:1, metal:1}: ‘containing or made of or resembling or characteristic of a metal’ are derived from the literal метал from the synset {метал:1, метален елемент} – {metallic element:1, metal:1}. However, the corresponding derived relation is not explicit in the Bulgarian wordnet. We can distinguish several cases of derivational implication: both literals are included in BulNet but the derivational relation between them is not expressed; either one or both derivationally related literals are not encoded yet in BulNet; language-specific derivational relations, such as diminutives, are not defined yet.

### 3.3 Language specific derivational relations

There are systematic morphosemantic differences concerning derivational mechanisms between English and Slavic languages (Koeva et al., 2008). Some of the most productive derivational relations in Bulgarian are briefly presented here, namely, verbal aspect pairs, gender pairs, and diminutives.

The Bulgarian verbs are classified as: imperfective (they express a process (duration, recurrence) or lack of integrity), perfective (they express integrity and completeness), bi-aspectual, imperfectiva tantum (if a perfective correspondent does not exist), perfectiva tantum (if an imperfective correspondent does not exist). Until now the aspect pairs have been introduced in one and the same synset in BulNet with an LNote describing the respective aspect. This representation is not sufficient, because perfect and imperfect verbs express clear difference in meaning reflected in their formal behavior: aspect verb pairs in Bulgarian have different lexicalization, belong to different word subclasses and to different inflection types, have different verb frames, different usage with respect to verb tenses and different derivatives, if any. In order to distinguish perfect and imperfect verbs, they are split into separate synsets subordinate to the same immediate hypernym (Figure 1) while the transitive hypernymy relation is based on imperfect verbs only. For example, the synset  $\{\text{скицирам:1, рисувам:1, нарисувам:1}\} - \{\text{draw:6}\}$ : 'represent by making a drawing of, as with a pencil, chalk, etc. on a surface' is split in  $\{\text{скицирам:1, рисувам:2}\} - \{\text{draw:6}\}$  and  $\{\text{нарисувам:1}\} - \{\text{paint on}\}$ . Derivationally related aspect verb pairs  $\{\text{рисувам:2}\}$  and  $\{\text{нарисувам:1}\}$  are linked with a literal relation, while the synsets are linked with a morphosemantic relation called 'aspect'.

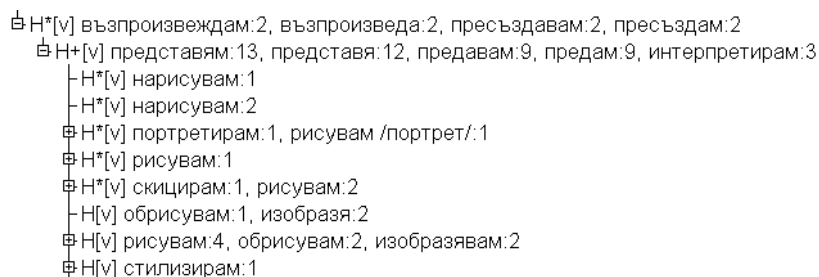


FIGURE 1: The aspect pairs are split in two synsets

The derivation of feminine nouns denoting an occupation from the respective masculine nouns is also productive in Bulgarian, although there are certain peculiarities that have to be taken into account. In some cases the feminine counterpart does not exist or is rarely used, in other cases when such feminine noun exists, the derived word does not mean quite the same or has acquired a completely different meaning. In the general cases, the feminine derivatives in Bulgarian are to be considered unique concepts as they have different lexicalization, belong to different word subclasses and to different inflexion types, have different usage, and are related to different hyponyms and derivatives, if any. Following the PWN practice, the feminine counterparts from the gender pairs were encoded in the BulNet

as hyponyms of the corresponding synset with the masculine counterpart (as a general rule, there is no corresponding concept lexicalized in English). The most appropriate solution is that the gender pairs be represented as sister hyponyms (if only gender difference is expressed) of the common hypernym. For example, the synset {фривзор:1, стилист:1} – {hairdresser:1, hairstylist:1, stylist:2, styler:1} has a hyponym {брџснар:1} – {barber:2}. The two synsets has feminine counterparts in Bulgarian: {фривзорка, стилистка} – 'female stylist' and {брџснарка} – 'female barber'. If the synset {фривзорка, стилистка} is presented as a hyponym of {фривзор:1, стилист:1}, it will have the same position in the wordnet structure as {брџснар:1}. The respective derivational relations between literals, e.g. {фривзор:1} – {фривзорка} and the morphosemantic relation between gender pair synsets are also linked.

The diminutives were included in BulNet only in the rare cases when the English equivalent is lexicalized. On the other hand, almost every concrete Bulgarian noun can have a diminutive (in some cases more than one). Thus, an appropriate position for the diminutives in the wordnet structure has to be provided. One option is to treat the diminutives in a way similar to the verb aspect and gender pairs. Another solution is to provide source literals with a derivation description encompassing the phenomena as well. The first approach will increase BulNet enormously; while the second one will treat the diminutives as a kind of word forms, even though most of them denote unique concepts. Moreover, there are diminutives that express only the emotional attitude and others that might be built according to the language rules but are almost never used. In general, the diminutives have different lexicalization with respect to the source noun, belong to different word subclasses and to different inflexion types, have different usage, and are related with different derivatives, if any. Taking into account all these considerations, there are more arguments for explicit mentioning of diminutives in the wordnet structure – as hyponyms with respect to to the source noun. e.g. {масичка:1, масичка за хранене:1}: 'a little table at which meals are served' is a hyponym of {трапеза:2, маса:8, маса за хранене:1} – {dining table:1, board:9}: 'a table at which meals are served'. The derivational relations between respective literals, e.g. {masa:8} and {masichka:1} are linked, as well as a morphosemantic relation 'diminutive' between synsets. The representation of aspect pairs, gender pairs and diminutives in BulNet linked with a hypernymy – hyponymy relation is subject to the entire organization of wordnet, where main nouns and verbs hierarchies are hypernymy – hyponymy.

#### 4 Nature of derivational relations

The nature of derivational relations requires that they be encoded as inverse and intransitive, so far as they usually affect various parts of speech and derivational mechanisms. Of course, relations such *асуча (study) – учител (teacher) – учителски (of a teacher)*, can be traced back. The difficulties in defining the derivational relations in wordnet come from the fact that polysemous words and homonyms are not distinguished from each other and, thus, one and the same derivational relation may be spread over all senses of a given literal, as well as

over all the senses of its derivative literal. Taking into consideration the polysemy, the following instances are observable in BulNet: monosemous source literal and monosemous derived literal; polysemous source literal and monosemous derived literal; monosemous source literal and polysemous derived literal; polysemous source literal and polysemous derived literal. The latter of these four cases gives grounds for the assumption that, at present, some of the derivational links (when explicit) are many-to-many, not taking into account the real semantic relations. In some cases, the use of such links might cause serious problems.

The automatic enrichment of wordnet on the basis of the derivational relations has been proposed and used recently for the Czech wordnet (Pala and Hlavachkova, 2007). If derivational morphology is applied automatically to a wordnet, regardless of whether there is also a semantic relation between the words, that will lead to the inclusion of a vast number of links that are, in reality, irrelevant, e.g. among other relations *мина* – *mine*: ‘explosive device that explodes on contact; designed to destroy vehicles or ships or to kill or maim personnel’ will be related with *минен* – *mining*: ‘the act of extracting ores or coal etc from the earth’. Therefore, derivational relations should be expressed only between literals that are semantically related (Miller and Fellbaum, 2003; Koeva et al., 2008). For example, the word *кристал* has several meanings {кристал:1} – {crystal:1}: ‘a solid formed by the solidification of a chemical and having a highly regular atomic structure’ and {кристал:1} – {crystal:5}: ‘glassware made of quartz’ among them. The derived relative adjectives have corresponding meanings related to their source nouns: {кристален:1}: ‘related to a crystal’ and {кристален:2}: ‘related to glassware made of quartz’. We might state that, in general, the derived words inherit the unique senses of their sources. We consider that only a limited couple of tasks may be done semi-automatically: suggesting literals linked by derivational relations instead of synsets; and identifying synsets where the potentially derivationally related literals appear.

A given derivational type can be seen as a model for word building. A given affix distinguishes the derived and source words and defines the affiliation of the derived word to a particular derivational type and word class. Words are classified in one and the same derivational type if a common derivational meaning is expressed by the affixes used. For example, the Bulgarian nouns *съседка*: ‘female neighbor’, *доставка*: ‘delivery’, *книжка*: ‘a small book’, do not constitute one derivational type although they are built with the same suffix *-к(а)*: the word *съседка* is built from a masculine noun, the word *доставка* – from a verb, and the word *книжка* is a diminutive. The derivational system of Bulgarian is represented by various types and subtypes.

## 5 Morphosemantic relations

Miller and Fellbaum (2003) describe the addition of “morphosemantic links” to WordNet that connect words (synset members) similar in meaning, in which one word is derived from the other by means of a morphological affix. Fellbaum et al. (2007) suggest that the meanings of affixes can be classified into a relatively small number of semantic categories, thus, the morphosemantic links can be labeled as

agent, instrument, etc. The Bulgarian linguistic tradition also acknowledges that the derivatives have not only a lexical and grammatical but also 'derivational' sense. That is the generalized sense of a series of derivatives formed from their sources with the same derivational affix. For example, the derivational sense of the words *nucameλ*: 'writer', *yчumeλ*: 'teacher', *чумameλ*: 'reader', *зoвoрyмeλ*: 'speaker' is 'a person performing the action, named by the source verb' and that sense is expressed by the suffix *-meλ*, hence, *чумameλ* is 'a person who reads', *nucameλ* – 'a person, who writes', etc.

It has been pointed out that a given morphosemantic relation may be expressed by different derivation mechanisms. On the other hand, different derivational mechanisms might indicate different semantic relations (Fellbaum et al., 2007). We might claim that morphosemantic relations are not language-specific, in contrast to the derivational mechanisms for lexicalisation. We can define the morphosemantic relation as a kind of semantic relation, indicated by a derivational relation in at least one language. The literal *paint* from the synset *{paint:3}*: 'make a painting of' has several derivatives participating in the synsets: *{paint:1}*: 'a substance used as a coating to protect or decorate a surface'; *{painter:1}*: 'an artist who paints'; *{painting:1, picture:2}*: 'graphic art consisting of an artistic composition made by applying paints to a surface'; *{painting:2}*: 'creating a picture with paints'. Neither of the corresponding Bulgarian synsets: *{боя:2}*, *{живописец:1, художник:1}*, *{картина:3}*, *{живопис:1}* consists a literal derived from *{писувам:2}*, the Bulgarian synset equivalent to *{paint:3}* (Koeva et al., 2008). Thus the same morphosemantic oppositions exist in Bulgarian, as well, roughly identified as a tool, agent, product, and activity, even though they are not derivationally marked. Derivational relations in certain language can be successfully employed for, but not limited to, the identification of a given morphosemantic opposition. Furthermore, they can be used for the identification of corresponding semantic relations in other languages, which have different means for lexicalization. If the lexical-semantic networks for a number of languages are connected, the semantically related synsets in one of the languages, for which relations are explicit through derivational relations, can be employed in order to connect to the same semantic relations in another language, which does not express those relations in terms of derivation. Therefore, a clear distinction should be made between derivation as relation between literals (inverse and intransitive) and the morphosemantic relations between synsets (possibly) explicitly marked by the derivation. Morphosemantic relations are a type of semantic relations that are language-independent. Morphosemantic relations form subclasses among the word classes: e.g. nouns that can act as human agents, nouns that can be act as inanimate agents, etc. Their exact definitions are to be developed on the basis of existing derivational relations in at least one language and, also, on the basis of their meaning.

## 6 Wordnet language external relations

Apart from the internal language relations, wordnet includes interlingual relations, as well, which connect equivalent synsets from the lexical-semantic networks of dif-

ferent languages according to a language independent inter-lingual-index. Wordnets include only concepts for which lexical expressions exist in the language (of course, every natural language has its means of expressing non-lexicalised concepts). Each wordnet is considered an individual and language-specific lexical-semantic network because the sets of lexicalized concepts are different for each language. That means that the hierarchical structure of the wordnet is different for the individual languages. In creating the BalkaNet, the Structure Preserving Principle was adhered to Tufiş et al. 2004, according to which the structure of English wordnet which has been taken originally as an inter-lingual index should be reflected in the other lexical-semantic networks. For the English synsets that do not have corresponding concepts or that were not lexicalized in the Balkan languages, the node was preserved and marked with the phrase 'no lexicalization'. The opposite trend is observable as well – lexicalized concepts in the Balkan languages that do not have lexical expressions in English. The productive derivational morphology of Slavic languages poses the question of establishing proper external relations between wordnets providing place for both language specific concepts and lexicalizations, while at the same time preserving correspondences.

Recently, an attempt has been made to establish a comprehensive worldwide wordnet Grid in which concepts will be stored lexicalized, without being dependent on a particular language, in order to overcome the approach in which wordnets are developed following the structure of PWN. The global wordnet grid consists of common base concepts (a set of concepts that play a major role in the building of wordnets), the base-level concepts (a set of concepts that are neither too general nor too specific), other concepts lexicalized in languages that relate to the first two sets (Fellbaum and Vossen, 2007). The global wordnet should include all concepts that have lexical expressions in at least one natural language. Therefore, the task of proper encoding of different levels of lexicalization in different languages is becoming more and more important in view of the various NLP tasks. Although the Slavic wordnets do not yet compete with PWN's coverage, they are continuously extended and improved so that a balanced global multilingual wordnet is foreseen.

## 7 Conclusion

We have briefly presented the current stage of the encoding of derivational relations in Bulgarian wordnet. We provided also some observations on the sophisticated nature of morphosemantic relations and presented some arguments that prove the negative consequences of purely automatic insertion of derivational relations into the wordnet structure. The further development of the Bulgarian wordnet is strictly connected to the investigation of the theoretical grounds of the nature of derivational and morphosemantic relations. At the first stage, the encoding of derivational relations between exact literals instead of synsets is foreseen. Another important task is the introduction of Bulgarian language-specific derivations in a uniform way providing at the same time interlingual correspondences. The existence of productive derivational models in languages such as Bulgarian (sharing many similar features of other Slavic languages) is a good starting point for the

gradual enrichment of the “global” wordnet – both in its connectivity and its lexical density references.

## References

- Christiane FELLBAUM, Anne Osherson, Peter E. Clark (2007), Putting Semantics into WordNet’s “Morphosemantic” Links, in: *Proceedings from 3rd Language and Technology Conference: HLT as a Challenge for Computer Science and Linguistics*, Poznan, Poland.
- Christiane FELLBAUM and Pick VOSSEN (2007), Connecting the Universal to the Specific: Towards the Global Grid, *Proceedings of The First International Workshop on Intercultural Collaboration (IWIC 2007)*, Kyoto, Japan.
- Svetla KOEVA, Tinko Tinchev and Stoyan Mihov (2004), Bulgarian Wordnet – Structure and Validation, in: *Romanian Journal on Information Science and Technology*, Vol. 7, 61-79.
- Svetla KOEVA, Cvetana Crstev and Dusko Vitas (2008), Morpho-semantic relations in WordNet – a case study for two Slavic Languages, in: *Proceedings of the 4th GWC*, 239–254.
- George MILLER and Christiane FELLBAUM (2003), Morphosemantic links in Wordnet, in: *Traitement automatique des langues*, 44.2:69–80.
- Karel PALA and D. HLAVACHKOVA (2003), Derivational Relations in Czech Wordnet, in: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, ACL, Prague, 75–81.
- Dan TUFIŞ, Dan Cristea and Sofia Stamou (2004), BalkaNet: Aims, Methods, Results and Perspectives., in: *Romanian Journal on Information Science and Technology*, Vol. 7, No. 1-2, 1-32.