

# Coping with Derivation in the Bulgarian Wordnet

**Tsvetana Dimitrova**

Institute for Bulgarian Language  
Bulgarian Academy of Sciences

cvetana@dcl.bas.bg

**Ekaterina Tarpomanova**

Faculty of Slavic Studies  
Sofia University

katja@dcl.bas.bg

**Borislav Rizov**

Institute for Bulgarian Language  
Bulgarian Academy of Sciences

boby@dcl.bas.bg

## Abstract

The paper motivates a strategy for identification and annotation of derivational relations in the Bulgarian wordnet that aims at coping with the complex morphology of the language in an elegant way. Our method involves transfer of the Princeton WordNet (morpho)semantic relations into the Bulgarian wordnet, at the level of the synset, and further detection of derivational relations between literals in Bulgarian. Derivational relations have been annotated to reflect the complexity of Bulgarian morphology. Introduced literal relations improve the consistency and employability of the wordnet.

## 1 Introduction

Bulgarian is a language with rich derivational morphology but derivational relations in the Bulgarian wordnet (BulNet) have been marked so far only at the level of the synset (Koeva, 2008). This paper outlines our strategy for representing the derivational relations at the level of the literal. We advocate for an approach with a twofold aim – to reflect the language specificity and to keep the overall structure of the Princeton WordNet (PWN) while modifying the representation of derivational and morphosemantic information (Fellbaum et al., 2009). We focus on noun-verb pairs (for encoding of other derivational patterns, cf. Koeva, 2008; Stoyanova et al., 2013). The derivational relations are to be further exploited for a prediction of (morpho)semantic relations between synsets that are not part of BulNet yet (hence, they are not found in PWN as they have no morphological realisation in English). As morphosemantic

relations in PWN are transferred into BulNet, we have used them to find prospective derivationally related pairs and derivational models in Bulgarian. Thus, the introduction of derivational relations improves connectivity in BulNet by explicitly linking morphological and semantic information through encoding links between literals in synsets connected via (morpho)semantic relations.

While encoding this information on different levels to reflect different phenomena, we enrich BulNet with information about derivational patterns that can be used in NLP tasks such as information retrieval and question answering (cf. Hathout and Tanguy, 2002; Ligozat et al., 2012).

In the next section, we briefly present the Bulgarian wordnet with some remarks on the specific conventions adopted for its development. Section 3 discusses other attempts at encoding derivational relations in wordnets of languages with rich morphology. The complexity of Bulgarian derivational morphology is outlined in Section 4. In Section 5, we brush on the first step of our method for automatic identification and annotation of derivational relations. Section 6 presents the set of conventions followed in the annotation of derivational relations that have been specified, along with the manual validation and correction of the results of the method applied (as introduced in Section 5). In Section 7, we outline directions for future work.

## 2 Bulgarian Wordnet – an Overview

The Bulgarian wordnet was launched as part of the BalkaNet project that aimed at creating a multilingual lexical database of wordnets for

Bulgarian, Greek, Romanian, Serbian, Turkish, and Czech (Stamou et al., 2002). BulNet aims to preserve the original structure of the Princeton WordNet and EuroWordNet (Vossen, 2004). Non-lexicalized synsets from PWN are kept in the overall structure and marked with the label *no lexicalization*. Language-specific concepts are included in the appropriate place of the lexical hierarchy.

Currently, BulNet comprises over 50,000 synsets. Unlike PWN which contains only open-class words, BulNet is enriched with function words (in synsets) added for the development of the Bulgarian Sense-Annotated Corpus, where every word is linked to a corresponding sense (synset) (Koeva et al., 2011). Words in BulNet are distributed into nine parts-of-speech: noun, verb, adjective, adverb, pronoun, preposition, conjunction, particle, and interjection – see the numbers in Table 1.

Part-of-speech	Count
Nouns	33,825
Verbs	6,199
Adjectives	8,114
Adverbs	1,395
Pronouns	94
Prepositions	423
Conjunctions	108
Particles	57
Interjections	11
Total	50,226

Table 1: Parts-of-speech distribution in BulNet

The main part of the relations in BulNet are semantic: *also\_see*, *causes*, *holo\_member*, *holo\_part*, *holo\_portion*, *hypernym*, *near\_antonym*, *similar\_to*, *subevent*, *verb\_group*. The list of semantic relations is based on the PWN lexical and conceptual relations (Koeva et al., 2004). BulNet encodes several morphosemantic – *be\_in\_state*, *bg\_derivative*, and morphological (derivational) relations – *derived*, *participle*. *Be\_in\_state* is a relation between an adjective and a noun considered as state of the respective adjective: {амбициозност:1, амбиция:1}<sup>1</sup> – {ambition:2, ambitiousness:1} is a state of {амбициозен:1} – {ambitious:1}. *Bg\_derivative* links a verb and a

noun derived from it that are semantically related, as in: {дирижирам:1} – {conduct:3} and {диригент:1, музикален ръководител:1} – {conductor:2, music director:1, director:1}. The relation *bg\_derivative* is transferred from PWN. *Derived* is a relation between a noun and an adjective derived from it, as in {каменен:1} 'made of, characteristic of or related to stone' derived from {камък:1} – {stone:4}. *Participle* is a relation between a verb and its participle – {пулверизиран:1} – {spray-dried:1} is a participle of {пулверизирам:1} – {spray-dry:1}.

### 3 Previous Work

Derivational relations in the Princeton WordNet 3.0 have been extracted through automatic identification of base-derived and semantically related noun-verb pairs (Fellbaum et al., 2009). A set of semantic relations across a number of morphologically derived noun-verb classes was determined, and morphological relations were added. The identified morphosemantic links connect word pairs where one of the literals is derived from another. They are marked as related both derivationally (relations *derived/derivative*) and semantically (relations *event*, *state*, *result*, *agent*, *undergoer*, *property*, *instrument*, *location*, *means*, *uses*, *destination*, *material*, *body part*, *vehicle*). Derivational pairs are available in a morphosemantic database through: <http://wordnet.princeton.edu/wordnet/download/standoff/>.

Other approaches involve automatic or semi-automatic adding of new synsets to wordnet by automatically deriving new words from already existing ones and adding morphological relations. Attempts at annotation of derivational relations are mostly made for wordnets of languages with rich morphology such as Romanian, Turkish, Estonian, and Slavic languages. Some approaches involve semi-automatic and automatic identification of derived word forms and pairs. The morphological analyser *Ajka* used for the Czech wordnet, works with a list of stems from which word forms are generated. A set of words is defined by identifying prefix, suffix, and a morphological tag, and a derivational rule is applied using a substitution of morphemes (affixes), with manual modification (deleting and correcting) of the generated word forms (Pala and Hlaváčková, 2007). The analyser *Derywator* is used for semi-automated expansion of the Polish wordnet through a combination of prefix and suffix modules in two transducers trained to

<sup>1</sup> Curly brackets mark a synset, and square brackets mark a literal.

work in the opposite direction on pairs already described in the wordnet and extended with automated construction of mappings representing internal stem alternations (Piasecki et al. 2012). For adding morphosemantic relations to the Romanian wordnet, simple literals were extracted (Mititelu, 2012). A list of prefixes and a list of suffixes were used to form combinations, and resulting forms are matched against a list of literals in the wordnet. Estonian wordnet was also enriched with synsets that are automatically generated using derivational suffixes (Kahusk et al., 2010).

Wordnets of other languages use language-specific labels and relations. Czech and Turkish wordnets adopt a set of labels that is different from the PWN ones. The Czech wordnet uses labels referring to part-of-speech: *deriv-na*, *deriv-dvrb*, *deriv-an*, *deriv-pos*, *deriv-pas*, *deriv-aad*, *deriv-an*, *deriv-g*, *deriv-ag*, *deriv-dem* (Pala and Hlaváčková, 2007). Labels in the Turkish wordnet are more general such as *become*, *acquire*, *be-in-state*, *something-with*, *someone-with*, *someone-from*, *someone-without*, *something-without*, *pertains-to*, *with*, *reciprocal*, *causes*, *is-caused-by*, *cat-of*, *manner* (Bilgin et al., 2004). The work on the Croatian wordnet (Katunar and Šojat, 2011) plans to follow the morphosemantic field model (Raffaelli and Kerovec, 2008).

Previous attempts at adding derivational relations to BulNet are outlined in (Koeva, 2008), (Koeva et al., 2008) and (Stoyanova et al., 2013). The derivational relations in PWN are transferred into and aligned to BulNet. They are marked at the level of the synset with *bg\_derivative* relation, or in *snote* when the transferred relation does not hold. Koeva (2008) proposes an approach for enlargement of BulNet that involves splitting verb synsets that contain both perfective and imperfective verbs.

The approach outlined in our paper involves automatic detection of candidate pairs and manual validation following language-specific conventions without straining too far from PWN. Morphosemantic relations hold among semantically related words sharing a stem with a close meaning. Semantic labels have been specified following PWN. After automatic detection of candidate pairs using the PWN morphosemantic database, we assign derivational relations to the identified literals. Next section gives a brief overview of some features of Bulgarian morphology to motivate our decisions for the annotation conventions adopted.

## 4 Bulgarian Derivational Morphology

Due to historical and cultural factors, Bulgarian language has preserved many Slavic features and acquired others that are common for the Balkan Sprachbund. Bulgarian is the only Slavic language with analytic nominal system, compensated by complex verb forms marked for aspect, mood, tense, voice, and evidentiality. Bulgarian derivational morphology combines inherited and borrowed word formants and shows great diversity of patterns.

### 4.1 Derivation Means

As in other Indo-European languages, there are two main morphological processes for formation of new words in Bulgarian – affixation and composition. Affixation consists in adding affixes to the root or the stem. Root is the central morpheme of a word that carries the main part of its semantic content, while stem is the root plus all derivational affixes, e.g., in *discounted* the root is *-count-* and the stem is *discount-*. Composition is defined as word formation by linking two or more stems. In Bulgarian, stems are often attached to each other by a linking morpheme – interfix, as in *вод-О-над* 'waterfall'.

A relatively rare word formation means is paradigmatic derivation – a term used to denote a derivation where the derivative keeps the same stem, but differs from the source word by its paradigm (Radeva, 1991: 51), as in *работя* 'to work' – *работа* 'work'. Paradigmatic derivation may occur in different directions: noun-to-noun, noun-to-verb, verb-to-noun, adjective-to-verb, etc. Inflection markers and/or thematic vowels may be added, removed, or replaced, as in: *десет* 'ten' – *десети* 'tenth'; *ниже* 'to string' – *низ* 'a string'; *тъга* 'a grief' – *тъжи* 'to grieve'. In this paper, we will use the term *conversion* to designate such instances of zero-suffixation.

### 4.2 Derivation Formants

Suffixation is the most productive derivational process in Bulgarian, and the most complicated one. Suffixes are polysemic, i.e., one suffix usually has more than one meaning. For instance, the suffix *-ник* is used to form nouns for agent (*проповядвам* 'preach' – *проповедник* 'preacher'), instrument (*подема* 'lift up' – *подемник* 'gig'), location (*багаж* 'luggage' – *багажник* 'luggage-carrier'), etc. The same meaning may be expressed by different suffixes such as the agentive *-ач*, *-ар*, *-ец*, *-ник*, *-тел*, *-ко*, *-льо*, *-ент/-ант*, *-атор*, *-джия/-чия*, etc.

In terms of origin, suffixes are domestic or borrowed from different languages – Turkish (-джия), Latin, directly or more often through intermediate language (-ция), Russian (-чик), English (-инг), etc. Some of the borrowed suffixes become productive and may be attached to domestic stems or even to stems borrowed from other languages, as the Turkish -джия/-чия in *таксиджия* 'taxi driver' and *интересчия* 'someone who is looking after his own interests'.

New words are formed by attaching one or more suffixes to the root or the stem. Suffixes may be added to the stem by agglutination to form a derivation chain, as in: *меля* 'to mill'; *мельница* 'a mill', where the suffix for location -нищ- is added to the verb stem; *мельничар* 'miller', with the suffix for agent -ар added to the noun stem; *мельничарски* 'characteristic or belonging to a miller', with the suffix for property -ск-.

Apart from agglutination, suffixation involves diachronic changes in the root or the stem, decomposition of the morphological structure, fusion between suffixes, between the suffix and the stem or between the suffix and the inflection, so that morpheme boundaries may become unclear. We will illustrate this process by two examples.

1) There are two possible analyses of the morpheme structure of imperfective verbs formed with the imperfectivating suffix -ва-, such as *зребвам* 'to scoop': *зреб-ва-м* (root – imperfectivating suffix – inflection marker), or *зреб-в-а-м* (root – imperfectivating suffix – thematic vowel – inflection marker). Both interpretations are possible (Ganeva, 2010: 135).

2) The words *летища* 'airports' and *сънища* 'dreams' seem to have the same derivational model. In fact, they have different morpheme structure: *лет-ищ-а* (root – suffix for location – inflection marker for plural) and *сън-ища* (root – inflection marker for plural). The paradigm of the second word is formed by analogy and was subjected to stem decomposition.

Unlike suffixes, prefixes do not cause any changes in the stem. Derivatives formed by prefixation do not change their part-of-speech. Prefixation is a typical means for verbal derivation that involves change of verbal aspect, namely perfectivation: *пиша* 'to write-impf<sup>2</sup>' – *напиша* 'to write-pf'. Polyprefixation is characteristic for Bulgarian, where every prefix modifies the se-

mantics of the word: *пиша* 'to write' - *пре-пиша* 'to copy out' - *до-пре-пиша* 'to copy out the rest'.

Both a prefix and a suffix can be attached to a stem to form a derivative, as in *вода* 'water' – *под-вод-ен* 'under-water'.

### 4.3 Phonetic Alternations

Derivation in Bulgarian is sometimes accompanied by phonetic changes that impede automatic detection of derivatives. Phonetic alternations are inherited from Old Bulgarian, and some of them are regular and still functional in Modern Bulgarian. Ablaut is a vowel alternation in the root that reflects word class or grammatical category, as in: *из-бИр-а-м* 'to choose' – verb, imperfective; *из-бЕр-а* 'to choose' – verb, perfective; *из-бр-ан* 'chosen' – participle; *из-бОр* – noun. Umlaut is a vowel alternation [a]/[e] depending on the stress and the vowel in the next syllable. It can express number, as in: *бЯл* 'white-m,sg', *бЯл-а* 'white-f,sg', *бЯл-о* 'white-n,sg' vs. *бЕл-и* 'white-pl'. Consonant alternations are due to historical palatalization and other phonetic laws. Some typical consonant alternations are: *к (κ)/ts (ц)/tch (ч)* – *човеК* 'man', *човеЧе* 'man-vocative', *човеЦи* 'men'; *т (m)/sht (щ)* – *свеТя* 'shine', *свеЩ* 'candle'; *д (d)/zhd (жд)* – *ограДа* 'enclosure', *ограЖДам* 'enclose'.

Some of the phonetic alternations have a grammatical value, but they are not considered derivational means.

### 4.4 Derivation vs. Inflection

In Bulgarian, inflection marks verbs for person, number, tense, voice, and mood, and nominal word classes – for gender and person (and case for pronouns). Inflection markers usually stand at the end of the word, after the derivational suffix(es), with the exception of some old word forms where an inflection may appear within the word (*м-О-ва* 'this-n,sg'), and before the definite article in nominal word forms (*жен-И-те* 'the women'). In our work, inflection markers are not taken into account as they affect only word forms and have grammatical meaning, in contrast to derivational affixes. Still, there are several grammatical suffixes in Bulgarian that have a contradictory interpretation.

Thematic vowels in Bulgarian are inherited from Proto-Slavic, and were further subjected to complex diachronic modifications. In Modern Bulgarian, thematic vowels are considered classificatory suffixes showing a verb conjugation and/or tense. Unlike derivational suffixes, they

<sup>2</sup> The following abbreviations are used in the paper: 'impf' - imperfective verb; 'pf' - perfective verb; 'impf. t.' - imperfectivum tantum; 'pf. t.' - perfectivum tantum; 'f.' - feminine; 'm' - masculine; 'n' - neuter; 1p, 2p, 3p - first, second and third person, respectively; sg - singular; pl - plural.

do not have any semantic content, but are involved in the derivation of verbs from nouns or adjectives, e.g., *мъка* 'pain' – *мъчиИш* 'to torment-2p,sg'<sup>3</sup>, *червен* 'red' – *червенЕе* 'to reddened-3p,sg'. Bulgarian linguistic literature defines this mode of derivation as paradigmatic (see Section 4.1.).

Verbal aspect in Bulgarian has two opposed interpretations: 1) aspectual pairs are grammatical forms of the same word; or 2) they are separate words as they show difference in meaning, verb frame, inflection type, and usage (Koeva, 2008: 363). We follow the second interpretation, i.e., to define aspect suffixes as derivational.

Participles are not explicitly classified for part-of-speech. As non-finite verb forms, they are traditionally considered a part of the verb paradigm, but their morphological formants are defined as derivational and not inflectional suffixes, as in *ходи-л* 'walked' where *-л* is a derivation suffix for aorist active participle with a zero inflection for masculine (for details on the grammar of the contemporary Bulgarian literary language, cf. Gramatika na savremenniya bulgarski knizhoven ezik. T. 2 Morfologiya., 1982).

## 5 Automatic Identification of Derivational Relations in BulNet

For automatic detection of derivational relations in BulNet, we employ the applicable information encoded in PWN. The method applied does not require any additional language resources, such as dictionaries or lists of affixes. The first step is to query for pairs of synsets linked via a morphosemantic relation in PWN. If a given pair of synsets has a corresponding pair in BulNet, we search for a pair of literals in the corresponding synsets with similar representation, and add a derivational relation to the literals found.

Two literals are similar if at least one of the following conditions holds:

1. One of the literals is included into the other, i.e., is substring of it. They are similar by inclusion.
2. The two literals in a pair have a long enough common prefix (as a string of symbols in the beginning of the word form). Its length has to be at least half the length of the shorter literal. Therefore, they are defined as similar by prefix.
3. The two literals have a Levenshtein distance smaller than a given value. The value is

<sup>3</sup> Thematic vowels are not visible in 1p, sg, present tense of verbs, so examples are in 2p and 3p.

calculated as the minimum number of: the length of the first literal, the length of the second literal; the absolute value of difference of the lengths of the two literals + a constant tolerance (2).

After calculating the similarity, we identify the differences between the words (literals) defined as relations: *prefix*, *suffix*, and *conversion*. If the literals match, the pair receives the relation *conversion*. If the two literals in the pair have the same beginning (defined as a string of symbols in the beginning of the word), the relation *prefix* cannot be attached. If the two literals have the same ending, the label *suffix* is excluded. If a relation of the type *prefix*, *suffix*, and *conversion* is not found, we compare the lengths of the common strings at the beginning and at the end. If the beginning is greater, we assign a *suffix* relation. A *prefix* relation is assigned if the ending is greater.

After the automatic assignment of derivational relations, manual validation was performed on all pairs found and annotation conventions were adopted in order to assure uniform and consistent approach to the morphological patterns in Bulgarian. We introduced two additional derivational relations – *deriv* (unspecified derivation) and *noun\_suffix/verb\_suffix* (substitution) to reflect specific processes and patterns (see section 6). Derivational relations were automatically assigned to literals denoting both members of the aspectual verb pairs, e.g., [*премахване:3*] 'disposal' received *without\_suffix* relation to both [*премахвам:3*] 'to dispose-impf' and [*премахна:3*] 'to dispose-pf'. However, the direct derivational relation links it only to the imperfective verb (*премахва-м* > *премахва-не*), so we remove the automatically assigned relation to the perfective verb. In the next section, we discuss the annotation conventions adopted.

## 6 Conventions for Annotation of Derivational Relations in BulNet

Literals pertaining to different synsets are derivationally linked via three asymmetrical (*suffix/without\_suffix*, *prefix/without\_prefix*, *noun\_suffix/verb\_suffix*) and two symmetrical (*conversion*, *deriv*) derivational relations attached to the literals. Synsets which contain these literals are linked via (morpho)semantic relations transferred from PWN. Numbers about the annotated literals are given in Table 2.

Derivational relation	Count
suffix/without_suffix	2,352
noun_suffix/verb_suffix	296
prefix/without_prefix	241
conversion	177
deriv	21

Table 2: Number of literals with a derivational relation assigned

Literals in BulNet can be linked via more than one derivational relations reflecting different patterns. Our aim is to find and represent the highly productive derivational patterns in order to trace other words that exhibit them and can be linked through respective (morpho)semantic relations (and assigned semantic labels). The noun literals are derivationally linked to one of the verb literals in a synset that contains both members of an aspect verb pair. If two literals in a synset are linked via a direct derivational relation, we do not assign an indirect one (although it may be a member of the corresponding synset). For instance, the noun [връщане:6] ‘return’ is linked to the verb [връщам се:1] ‘to return’, and the noun with the prefix *за-* – [завръщане:3] ‘return’ is linked to the verb [завръщам се:1] ‘to return’ (respective literals are members of the same synsets – a noun and a verb one, respectively). However, there may be not a direct link, and we may link the two literals via an indirect derivational relation – we can observe further which pattern is more productive. The labels of derivational relations assigned do not reflect the real direction of the derivation. In the subsections, we will discuss the types of derivational relations assigned to verb-noun pairs.

### 6.1 Suffixation: suffix/without suffix

The derivational relation *suffix/without\_suffix* is asymmetrical and marks suffixation (when a suffix or a combination of suffixes are used to generate new words) and suffix removal, respectively, as in [плувам:1] ‘to swim’ / [плуване:1] ‘swimming’ where the deverbial noun suffix *-не* is attached to the stem of the verb *плува-* (*-м* is the inflection marker for 1p, sg, present form of the verb).

In BulNet, verbs are classified as imperfective, perfective, bi-aspectual, imperfectiva tantum, and perfectiva tantum (Коева, 2008). Though verbs in aspect pairs are members of one synset, they express difference in meaning, and

form different derivatives<sup>4</sup>. Deverbial nouns with suffix *-не* are derived from the imperfective stem and usually denote a process. Nouns ending in *-не* are derivationally linked to the literals of imperfective verbs. Deverbial nouns formed with the suffix *-ние* are derived from the aorist stem, usually denote a result of an action, and can be derivationally linked to perfective or imperfective verbs. The synset {миграция:1, мигриране:1, преселване:1, преселение:1} – {migration:1} ‘the movement of persons from one country or locality to another’ is linked as *event* to the synset {преселвам се:2, преселя се:2, мигрирам:1, разселвам се:1} – {migrate:1, transmigrate:1} ‘move from one country or region to another and settle there’. Literals are derivationally linked as follows:

```
{преселвам се:2, преселя се:2, мигрирам:1,
разселвам се:1}
has_event: {миграция:1, мигриране:1, преселване:1, преселение:1}
[преселвам се:2]
Inote: impf.
suffix: [преселване:1]
[преселя се:2]
Inote: pf.
suffix: [преселение:1]
[мигрирам:1]
Inote: impf. and pf.
suffix: [мигриране:1]
noun_suffix: [миграция:1]
```

A *-ние* noun can be derivationally linked to imperfectiva tantum verbs, such as: [тълкувам:2] – [interpret:3] ‘give an interpretation or explanation to’ and [тълкувание:1] and [тълкуване:2] (belonging to the same synset) – [interpretation:3] ‘a mental representation of the meaning or significance of something’.

In Bulgarian, participles can have both verbal interpretation (as in passive voice) and nominal one. If a participle is substantivised, i.e., is a member of a noun synset, and this synset is linked via a (morpho)semantic relation to a verb synset, the participle may receive a derivational relation. *Разлято* and *разляно* ‘spilled’ are both passive participles of the verb *разляя* ‘to spill’. Thus, {разлято:1, разляно:1} – {spill:1} ‘liquid that is spilled’ is an *event* of {разливам:1,

<sup>4</sup> The aspect pairs are introduced in one and the same synset (the aspect is mentioned in an Inote) to keep the symmetry with PWN. However, as this representation is not sufficient, they are to be split into separate synsets subordinate to the same immediate hypernym (Коева, 2008: 363).

*разляя:1, изливам:4, изляя:4, разсипвам:4, разсипя:4, изсипвам:1, изсипя:1* – {*spill:7, slop:2, splatter:2*} 'cause or allow (a liquid substance) to run or flow from a container' that have the following derivational relations:

[*разляя:1*]  
 Inote: pf.  
 suffix: [*разляно:1*]  
 suffix: [*разлято:1*]

The *-не* and *-ние* patterns are among the most productive. Most *-не* and *-ние* nouns in BulNet are members of synsets linked to the verbs via an *event* (morpho)semantic relation (1,207 of the synsets with *-не* nouns, and 448 with *-ние* nouns). 57 of the synsets containing *-ние* nouns and 43 of the *-не* nouns are linked to the verbs via *result* semantic relation. The *state* relation connects 42 of the synsets with *-не* nouns and 67 of the synsets of *-ние* nouns.

In order to find productive derivational patterns in Bulgarian, we mark derivational relations on literals that are indirectly related to the derivative (derived by another member of the chain) and show a pattern containing more than one suffix. The noun [*ковачница:1*] 'forge' is linked as *location* and via *suffix* to [*кова:2*] 'to forge' although *ковачница* is derived via *ковач* 'blacksmith' (PWN shows no derivational or (morpho)semantic relation between [*forge:5*] and [*blacksmith:1*]). The semantic relation between *кова* and *ковачница* is derivationally motivated – *forge* is a *location* where a blacksmith forges. The derivation path (verb + suffix for agent + suffix for location) may be applied to find other pairs with similar morphosemantic relation, as in *тъка* 'to weave' – *тъкач* 'weaver' – *тъкачница* 'weaving workshop'.

### 6.2 Substitution: *noun\_suffix/verb\_suffix*

The relation *noun\_suffix/verb\_suffix* is asymmetrical and marks a suffix on both members of the pair, as in [*аккомпанирам:1*] 'to accompany' and [*аккомпанимент:1*] 'accompaniment' – the suffix on the verb is *-ира-* and the noun suffix is *-(и)мент*. The derivation process involves two operations – removing a verb suffix and adding a noun suffix to form a noun and vice versa.

A literal can have several derivatives pertaining to the same or different synsets, as in [*епилирам:1*] – [*epilate:1*] 'remove body hair' linked via *suffix* relation to [*епилиране:1*] – [*epilation:1*], and via *noun\_suffix* relation to [*епиляция:1*] – both are *event* members of the

synset {*епилиране:1, епиляция:1, депилиране:1, депиляция:1, обезкосмяване:1*} – {*epilation:1, depilation:1*} 'the act of removing hair (as from an animal skin)'; and via *noun\_suffix* relation to material [*епилатор:1*] – [*epilator:1*] of the synset {*депилатор:1, депилатоар:1, епилатор:1*} – {*depilatory:2, depilator:1, epilator:1*} 'a cosmetic for temporary removal of undesired hair'.

### 6.3 Prefixation: *prefix/without\_prefix*

Another asymmetrical relation marks prefixation and prefix removal. In Bulgarian, prefixation does not change the part-of-speech, so adding or removing a prefix in noun-verb pairs is always accompanied by attachment of a thematic vowel to form a verb and its removal to form a noun, e.g., [*завинтя:1*] 'to screw' *without\_prefix* [*винт:1*] 'screw'. As thematic vowels do not have any semantic content, their attachment or removal is not explicitly annotated.

The relation *prefix/without\_prefix* can be combined with *suffix/without\_suffix* or *noun\_suffix/verb\_suffix* when the suffix has a lexical content as in *въоръжа* 'to arm' vs. *оръжие* 'armament' where the verb is derived via prefixation (prefix *въ-*) and the noun is derived via suffixation (suffix *-ие*). Thus, the synset {*въоръжа:1, въоръжавам:1*} – {*arm:2*} is related via the (morpho)semantic relation *uses* with the synset {*оръжие:1, въоръжение:1*} – {*armament:2*}, and the the literal [*оръжие:1*] is derivationally related to [*въоръжа:1*] via the relations *prefix* and *without\_suffix*.

{*оръжие:1, въоръжение:1*]  
 is\_used\_to: {*въоръжа:1, въоръжавам:1*]  
 [*оръжие:1*]  
 prefix: [*въоръжа:1*]  
 without\_suffix: [*въоръжа:1*]

Derivationally related verb-noun pairs via prefixation are much rarer – 241 instances (2,352 of suffixation).

### 6.4 Conversion

The symmetrical relation *conversion* (marked on both literals of the pair) annotates zero-suffixation, as in [*викам:1*] 'to cry' and [*вик:1*] 'a cry' – the thematic vowel *-а-* and the inflectional suffix for 1p, sg, present tense *-м* are removed and no derivational suffix is added to generate the noun. The reverse process of adding a thematic vowel and an inflection marker to form a verb, is also marked as *conversion*, e.g.,

[*посреднича:1*] ‘to mediate’ is derived by conversion from [*посредник:1*] ‘mediator’.

Derivational relations may link literals of the same synset to literal from different synsets:

{*тъжа:1, тъгувам:2, жалья:1*} – {*sorrow:1, grieve:1*} ‘feel grief’

has\_state: {*тъга:1, печал:2, униние:1*} – {*sorrow:5, sadness:3, sorrowfulness:2*} ‘the state of being sad’

has\_event: {*жал:1, мъка:3, печал:1*} – {*sorrow:3*} ‘an emotion of great sadness associated with loss or bereavement’

[*тъжа:1*]

Inote: impf. t.

conversion: [*тъга:1*]

[*жалья:1*]

Inote: impf. t.

conversion: [*жал:1*]

### 6.5 Not Otherwise Specified: *deriv*

The symmetrical relation *deriv* (derivative) marks both members of the pair if a derivational pattern is unclear, as in [*помогна:1*] ‘to help-pf’ / [*помагам:1*] ‘to help-impf’ and [*помощ:1*] ‘help’ – historically, *помощ* is a deverbal noun but the derivation is not transparent in modern Bulgarian.

We do not expect literals with a *deriv* relation to show evidence for any productive pattern.

## 7 Conclusion and Future Work

In this paper, we presented the first results of an approach for introduction of derivational relations into the Bulgarian wordnet. We discussed the specifics of the Bulgarian morphology to support the conventions adopted for annotation of derivational patterns in Bulgarian. We identified (automatically) and annotated (through automatic identification and assignment of derivational labels with manual validation and modification afterwards) a set of noun-verb pairs in the Bulgarian wordnet.

The work on annotation allows for an observation on derivational patterns that can be used to improve the process of automatic identification and assignment of relations (derivational and (morpho)semantic ones). For instance, the nouns with suffix *-(a/u)ция* denote: *event* (312 instances), *result* (46), *means* (28), *state* (17), *undergoer* (17), *uses* (16), *agent* (5).

The annotation will allow us to enrich the Bulgarian wordnet with new relations. In addition,

we can easily identify synsets that have not been created yet.

In the next stages of the experiment, we plan to rerun the automatic identification of derivational relations exploiting the newly specified relations/conventions. We can automatically detect derivational pairs using the patterns identified and link them with semantic relations. Automatic assignment of (morpho)semantic relations is also a potential direction to be exploited.

## Acknowledgments

The present paper was prepared within the project Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics (BG051PO001-3.3.06-0022) implemented with the financial support of the Human Resources Development Operational Programme 2007-2013 co-financed by the European Social Fund of the European Union. The authors take full responsibility for the content of the present paper and under no conditions can the conclusions made in it be considered an official position of the European Union or the Ministry of Education, Youth and Science of the Republic of Bulgaria.

## References

- Orhan Bilgin, Özlem Çetinoğlu, and Kemal Oflazer. 2004. Morphosemantic Relations In and Across Wordnets – A Study Based on Turkish. In *Proceedings of the Second Global Wordnet Conference*, pages 60–66.
- Christine Fellbaum, Anne Osherson, and Peter E. Clark. 2009. Putting Semantics into WordNet’s “Morphosemantic” Links. In *Proceedings of the Third Language and Technology Conference, Poznan, Poland*. Reprinted in: *Responding to Information Society Challenges: New Advances in Human Language Technologies*, Springer Lecture Notes in Informatics, vol. 5603, pages 350–358.
- Gergana Ganeva. 2010. Kam istoriyata na sufiksita za imperfektivacija v balgarskite dialekti. / On the History of the Imperfectivating Suffixes in Bulgarian Dialects. *Eslavística Complutense* 10. Madrid, pages 135–145.
- Gramatika na savremenniya balagrski knizhoven ezik. T. 2 Morfologiya*. Sofia: Izdatelstvo na Balgarskata akademiya na naukite. 1982. / Grammar of Contemporary Bulgarian Literary Language. Vol. 2, Morphology. Sofia: Bulgarian Academy of Sciences Publishing House. 1982.
- Nabil Hathout and Ludovic Tanguy. 2002. Webaffix: Discovering Morphological Links on the WWW.



- In *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Espagne, pages 1799–1804.
- Neeme Kahusk, Kadri Kerner, and Kadri Vider. 2010. Enriching Estonian Wordnet with Derivations and Semantic Relations. In *Human Language Technologies – the Baltic Perspective. Proceedings of the Fourth International Conference Baltic HLT 2010*. IOS Press, pages 195–200.
- Daniela Katunar and Krešimir Šojat. 2011 Morphosemantic fields in the building of the Croatian WordNet: the verbs of movement. In *Space in Time and Language*. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang GmbH, pages 79–89.
- Svetla Koeva, Tinko Tinchev, and Stoyan Mihov. 2004. Bulgarian Wordnet – Structure and Validation. *Romanian Journal of Information Science and Technology*, Vol. 7, No. 1-2, pages 61–78.
- Svetla Koeva. 2008. Derivational and Morphosemantic Relations in Bulgarian Wordnet. *Intelligent Information Systems*, XVI, Warsaw, Academic Publishing House, pages 359–389.
- Svetla Koeva, Cvetana Krstev, and Duško Vitas. 2008. Morpho-Semantic Relations in Wordnet - a Case Study for Two Slavic Languages. In *Proceedings of the Fourth Global WordNet Conference*, Szeged, pages 239–254.
- Svetla Koeva, Svetlozara Leseva, Borislav Rizov, Ekaterina Tarpomanova, Tsvetana Dimitrova, Hristina Kukova, and Maria Todorova. 2011. Design and Development of the Bulgarian Sense-Annotated Corpus. In *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de corpora. Actas del III Congreso Internacional de Lingüística de Corpus*. Valencia: Universitat Politècnica de València, pages 143–150.
- Anne-Laure Ligozat, Birgitte Grau, and Delphine Tribout. 2012. Morphological Resources for Precise Information Retrieval. In *Text, Speech and Dialogue. Proceedings of the 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012*. Lecture Notes in Computer Science, Volume 7499, pages 689–696.
- Verginica Barbu Mititelu. 2012. Adding Morpho-Semantic Relations to the Romanian Wordnet. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2596–2601.
- Karel Pala and Dana Hlaváčková. 2007. Derivational relations in Czech Wordnet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 75–81.
- Maciej Piasecki, Radosław Ramocki, and Marek Mażarz. 2012. Recognition of Polish Derivational Relations Based on Supervised Learning Scheme. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, European Language Resources Association (ELRA), pages 916–922.
- Vasilka Radeva. 1991. *Sloobrazuvaneto v balgarskiya knizhoven ezik*. Sofia: Universitetsko izdatelstvo Sv. Kliment Ohridski. / Word Formation in Bulgarian Language. Sofia: Sofia University Press.
- Ida Raffaelli and Barbara Kerovec. 2008. Morphosemantic fields in the analysis of Croatian vocabulary. *Jezikoslovlje* 9.1–2: 141–169.
- Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis, Svetla Koeva, George Totkov, Dominique Dutoit, and Maria Grigoriadou. 2002. BALKANET: A Multilingual Semantic Network for the Balkan Languages. In *Proceedings of the International Wordnet Conference, Mysore, India*, pages 21–25.
- Ivelina Stoyanova, Svetla Koeva, and Svetlozara Leseva. 2013. Wordnet-based Cross-Language Identification of Semantic Relations. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, Sofia, Bulgaria, 8-9 August 2013, pages 119–128.
- Piek Vossen. 2004. EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual Index. *International Journal of Lexicography*, 17(1): 161–173.