Papers of BAS
Humanities and Social Sciences
Vol. 8, 2021, No. 1

The Bulgarian WordNet: Structure and specific features

Svetla Koeva

Abstract. WordNet is a semantic network whose nodes host synsets denoting different concepts and whose arcs, connecting the nodes, encode different types of relations (semantic, morpho-semantic, derivational, extralinguistic). The current paper offers a short overview of the way the Bulgarian WordNet has been constructed with a special focus on the additional information it contains in comparison with the Princeton WordNet. The Bulgarian WordNet embraces in its structure words and multiword expressions from all parts of speech, it possesses an elaborated system of grammatical, lexical and stylistic labels describing the synonyms; it owns a detailed system of semantic classes: nouns are grouped in 256 classes, verbs – in 15 classes (as the Princeton WordNet), adjectives – in 17 classes, and adverbs – in 9 classes; the Bulgarian WordNet is expanded with derivational relations between nouns and verbs and corresponding morpho-semantic relations and with syntagmatic relations by means of Conceptual frames linking sets of verbs with sets of semantically and syntactically compatible nouns.

Key words: wordnet, Bulgarian WordNet, semantic relations

1. Introduction

WordNet is a semantic network whose nodes host synonyms denoting different concepts and whose arcs, connecting the nodes, encode different types of relations (semantic: genus-kind, part-whole, cause-effect, etc.; morpho-semantic: agent-predicate, predicate-instrument, predicate-state, etc.; derivational; extralinguistic, i.e., membership to a thematic domain; inter-language, i.e., translation equivalents). The idea for organizing the lexicon of a given language into a (lexico-)semantic network was first executed in the Princeton WordNet[1] (Miller et al. 1990/1993). Some of the fundamental ideas on which the WordNet is based encompass: a) the use of a semantic network which embraces taxono-

[1] http://wordnet.princeton.edu/

mies, meronomies and non-hierarchical relations with clearly defined properties which allow for quick and easy automatic processing; b) a different organization of the lexicon in comparison to traditional dictionaries where words are ordered alphabetically and the links among semantically related words (such as between sister hyponyms, between a whole and its parts, etc.) are not explicitly presented (Miller 1986).

For example, the nouns {airplane; aeroplane; plane} constitute a synonymous set (synset), which is described by the definition in the Princeton WordNet 'an aircraft that has a fixed wing and is powered by propellers or jets'[2]. This synset is connected to the more general synset – its hypernym: {heavier-than-air craft} 'a non-buoyant aircraft that requires a source of power to hold it aloft and to propel it'; with a number of more specific synsets – hyponyms: {bomber} 'a military aircraft that drops bombs during flight' and {fighter; fighter aircraft; attack aircraft} 'a high-speed military or naval airplane designed to destroy enemy aircraft in the air'; with synsets denoting parts of a whole – meronyms: {wing}; and with equivalent synsets from other languages: {самолет, аероплан} in Bulgarian, {αεροπλάνο} in Greek, {aeroplan; avion} in Romanian, {avion} in Serbian, {uçak; tayyare} in Turkish, etc.

The Bulgarian WordNet (BulNet)[3] has been and is being constructed since 2001 within various international and national projects by the team of the Department of Computational Linguistics at the Institute for the Bulgarian Language at the Bulgarian Academy of Sciences[4]. The paper offers a short overview of the way the Bulgarian WordNet is being developed with a special focus on the additional information it contains in comparison with the Princeton WordNet. The structure of the paper is as follows: in Section 2 the general structure of the Bulgarian WordNet is presented emphasizing that the Bulgarian WordNet embraces words and multiword expressions from all parts of speech; Section 3 introduces the ways of development of the Bulgarian WordNet; Section 4 is focused on upgrading the synset structure (in particular, on the system of grammatical, lexical and stylistic labels describing the synonyms; the structure of definitions; the extended ontology of semantic classes of nouns, adjectives and adverbs; the attribution of Conceptual frames to verb synsets); and in Section 5 the different types of relations linking synsets and literals in Bulgarian WordNet are described.

2. General overview of the WordNet structure

The synsets in the Princeton WordNet include synonyms (called literals), belonging to four parts of speech: noun, adjective, verb and adverb (cardinal and ordinal numerals are classified as adjectives, and some numerals with substantive use are classified as nouns; only some indefinite and negative pronouns

[2] The synsets, their definitions, and the remaining related information comes from either the Bulgarian WordNet or the Princeton WordNet and where appropriate these have been adapted by the author.

[3] http://dcl.bas.bg/bulnet/

[4] http://dcl.bas.bg

are included and they are determined either as adjectives or as adverbs in accordance with their characteristics).

"Nouns are organized as topical hierarchies, verbs are organized by a variety of entailment relations, and adjectives and adverbs are organized as N-dimensional hyperspaces" (Miller et al. 1990/1993, 3). Nouns and verbs build hierarchical structures, based on the relations of **hypernymy** and **hyponymy** (**troponymy** in the case of verbs) among the synsets. Descriptive adjectives are organized in structures which have "central" synsets, connected via the relation of antonymy, and satellite synsets, linked through the relations **similar to** and **also see** (Fellbaum, Gross, Miller 1990/1993, 28). A similar design is applied for adverbs, but many adverbs in the Princeton WordNet are not related with other synsets. Relational and participial adjectives do not follow the typical star-like structure; rather they are linked via derivational relations with a noun or verb respectively. Other semantic, derivational, and morpho-semantic relations are established between the noun, verb, adjective and adverb synsets, if such relations exist between the entities, events and properties denoted by the synsets. In practice, for the four part-of-speech classes four different (lexico-)semantic networks have been created among which extralinguistic and/or derivational and morpho-semantic relations (built upon the semantics of derivational relations) are established.

The synsets in the Bulgarian WordNet are divided into nine part-of-speech classes – pronouns, prepositions, conjunctions, particles and interjections have been added to nouns, verbs, adjectives and adverbs. The organization of nouns, verbs and adjectives in the Bulgarian WordNet follows the structure of the Princeton WordNet. All adverbs in the Bulgarian WordNet are linked through the extralinguistic relation **domain category** depending on their most general categorical meaning (for example, the synset {бързо; скоростно; светкавично} (barzo; skorostno; svetkavichno) 'quickly; rapidly' is linked to the synset {начин; стил; маниер} (nachin; stil; manier) 'manner' with the relation **domain category**).

In the case of numerals, the classification in the Princeton WordNet is applied (as nouns or adjectives), since this fully corresponds to their functional characteristics in Bulgarian. A few language specific numerals have been added, expressing approximation, for example {тридесетина; трийсетина}(tridesetina; triysetina) 'around thirty; about thirty'.

We share the view that the nature of open and closed word classes is different and consequently their organization in the WordNet structure should be different. For this reason, the synsets representing closed word classes are linked to the WordNet structure by means of extralinguistic relations.

The pronouns in Bulgarian stand for nouns, adjectives and adverbs (as well as for prepositional phrases with adverbial meaning). The members of the Bulgarian pronominal system, which are not included in the Princeton WordNet, are classified with the part-of-speech label pronoun, their use as nouns, adjectives or adverbs is pointed out in the definition and their English translation is included in a synset note. An additional classification is provided through the extralinguistic relation **domain category** with which the respective pronoun is related to the synset representing the type of pronoun: the pronouns' types are

introduced as hyponyms of the synset {pronoun}. For example, the pronoun {някой} (nyakoy) 'someone' with a definition 'indefinite pronoun (standing for a noun phrase) used to denote unknown person' is linked with the relation **domain category** to the synset {неопределително местоимение} (neopredelitelno mestoimenie) 'indefinite pronoun' and is translated in English as *someone*, *somebody* within a synset note.

In the Bulgarian WordNet all prepositions, conjunctions, particles and some interjections are also introduced, following the same conventions: the English equivalent (in case there is one) is provided in a synset note and the synsets are linked with the extralinguistic relation **domain category** to the synset which expresses their type. For example, the conjunction {kamo} (kato) 'as' with the definition 'subordinating conjunction introducing a subordinate adverbial clause' is linked with the relation **domain category** to the synset {подчинителен съюз} (podchinitelen sayuz) 'subordinating conjunction' with the hypernym {съюз} (sayuz) 'conjunction', while the preposition {kamo} (kato) 'like' with the definition 'preposition, introducing a transient or characterising feature, function, role, status, etc., which someone or something has or acts in accordance with, exists or presents themselves/itself as' is linked with the relation **domain category** to the synset {предлог} (predlog) 'preposition'.

In this way the Bulgarian WordNet has been expanded with new synsets for pronouns (94 synsets), prepositions (423 synsets), conjunctions (108 synsets), particles (57 synsets) and interjections (11 synsets).

Each synset is associated with a definition and usage examples and the synonyms are marked with an index in order to differentiate among polysemous and homonymous literals, which belong to different synsets.

Two types of semantic relations are included in the Princeton (and in the Bulgarian) WordNet: among literals (called lexical) and among synsets (called conceptual). The conceptual relations apply to all literals in both connected synsets, while the lexical relations link particular literals within one or two synsets. The conceptual relations among synsets are: **hypernymy** (*elephant* and *proboscidean*), including hypernymy to a proper noun, called **instant hypernymy; hyponymy** (*proboscidean* and *elephant*), including hyponymy of a proper noun, called **instant hyponymy; holonymy** (**holo part**, **holo substance**, **holo member**) (*finger* and *nail*, *bread* and *flour*, *basketball league* and *basketball team*); **meronymy** (**mero part**, **mero substance**, **mero member**) (*nail* and *finger*, *flour* and *bread*, *basketball team* and *basketball league*); **entailment** (**has subevent** and **is subevent of**) (*snore* and *sleep*); **causes** and **is caused by** (*kill* and *die*); **has attribute** and **is attribute of** (*posture* and *postural* 'of or relating to or involving posture'), **verb group** (*run* 'move fast by using one's feet, with one foot off the ground at any given time' and *run* 'cover by running; run a certain distance'), **similar to** (*divided* and *segmental*), **also see** (*complex* and *difficult*). The lexical relations linking literals are semantic: **synonymy** (*small* and *little*) and **antonymy** (*small* and *large*), and derivational: the general relation **derivationally related form** (*reader* and *read*), and the specific relations **has participle** and **is participle of** (*hammer* and *hammered*) and **has pertainym** and **is pertainym of** (*finance* and *financial*). The lexical relations constitute a different semantic network within the WordNet, where nodes are literals, not synsets. Extralinguistic relations have

also been introduced between synsets: **domain category** (*cloudy* and *meteorology*); **domain region** (*safari* and *Africa*); **domain usage** (*hammered* and *slang*). Each synset contains at least one literal and for each synset at least one intralingual relation to another synset has to be defined.

3. Development of the Bulgarian WordNet

Two basic approaches are typically used in the creation of (lexico-)semantic networks: the expand-model and the merge-model (Vossen 1996, 716), and a combination of the two. The first approach consists in the translation of the synsets, their definitions and usage examples from one language (most frequently from the Princeton WordNet) into another, wherein all relations between synsets, as well as the remaining semantic information, for example the semantic class of the synset, the sentence frames of verb synsets, etc., are transferred over (and manually or semi-automatically checked). The second approach involves an independent development of a wordnet for a respective language on the basis of existing language resources (monolingual dictionaries, dictionaries of synonyms and of antonyms, corpora, etc.) and its subsequent connection to other (lexico-)semantic networks. An example of such a wordnet is the Polish WordNet, which was created as an independent semantic network, after which it was connected to the Princeton WordNet (Rudnicka et al. 2012). The independent development of wordnets faces the problem of achieving sufficient overlap in the lexical coverage while maintaining language specific properties.

The Bulgarian WordNet has been created mainly by expansion of the Princeton WordNet, even though a number of automated procedures have been used for merging existing language resources at the level of synsets: synonyms from synonymous dictionaries, definitions from explanatory dictionaries and illustrative examples from dictionaries and corpora (Koeva, Totkov, Genov 2004, 46-49). With the development of technology, at present, the construction of the Bulgarian WordNet is based on a prior automatic translation of the synsets from the Princeton WordNet. The synsets, resulting from the automatic translation, are manually evaluated as follows: a) literals that do not express the meaning of the synsest are removed; b) literals that signify the meaning of the synset but have been represented in an inappropriate form (i.e., with spelling errors or with omitted or redundant parts) are edited; c) new literals that correspond to the meaning of the synset are included. The synset relations are automatically transferred from the Princeton WordNet and are then manually evaluated, whereby the approach to the changes (removal, addition, transformation) of relations is conservative. As a result of the automatic translation, the number of synsets in the Bulgarian WordNet is roughly equal to their number in the Princeton WordNet (Table 1). In comparison, the Princeton WordNet contains 117,659 synsets, which embrace 155,287 unique literals. More than two-thirds of the synsets in the Bulgarian WordNet have been evaluated and edited by experts. The reliability of the data in the Bulgarian WordNet is guaranteed in a number of ways – by the corresponding information in English (and possibly in other languages), by cross-references with existing dictionaries and corpora, and by manual evaluation.

Table 1. Quantitative data on the Bulgarian WordNet as of March 2021

Total number of synsets	121,282
Synsets created or evaluated by experts	85,954
Total number of synonyms	256,726
Synonyms created or evaluated by experts	168,433

The Princeton WordNet contains only concepts that have been lexicalized (admittedly, each language has the means to express non-lexicalized concepts as well). Each wordnet can be viewed as an independent, language-specific (lexico-)semantic network, since the set of lexicalized concepts is different for each language, respectively the structure of the wordnets is (or may be) different. In the building of the Bulgarian WordNet the **Principle of structure preservation** (Tufiş, Cristea, Stamou 2004, 15) is observed, according to which the structure of the Princeton WordNet has to be mirrored. The node for an English synset for which the concept has not been lexicalized in Bulgarian is preserved, the literal is presented descriptively and the definition is marked with the expression 'no lexicalization'. For example, the English literal {hot cereal} is not lexicalized in Bulgarian and is translated as {топла зърнена закуска} (topla zarnena zakuska) with the definition 'no lexicalization – a cereal that is served hot for breakfast'. The empty nodes in the Bulgarian WordNet serve not only to preserve the structure of the Princeton WordNet, but also to maintain the hierarchical organization of the taxonomic relations such as **hypernymy** and **hypomyny**.

The Princeton WordNet is the primary source for an abstract representation of an Inter-Lingual Index of meanings to which concepts for other languages are mapped (Vossen et al. 1998, 9). Each synset in the Bulgarian WordNet is labelled with an identification number, corresponding to an Inter-Lingual Index. The lack of a lexicalization in English is indicated within the structure of the Bulgarian WordNet as follows: a) the language specific Bulgarian synsets are related with unique identification numbers not present at the Princeton WordNet; the language specific literals are transliterated into Latin alphabet within literal notes and the synset definition is translated into English within a synset note; b) the relations with other synsets show the position of the language specific synset within the semantic network and through these relations the position of the corresponding empty node in the English WordNet can be identified. For example, the Bulgarian synset {зелева сарма} with a literal note: the transliteration *zeleva sarma*; and a synset note: the definition 'sarma rolled in a whole cabbage leaf, that has been pickled in brine', represents a language specific concept that does not exist in the Princeton WordNet, the same stands for its hypernym {сарма} with the literal note: the transliteration *sarma*, and the synset note: the definition 'a kind of dish made of minced meat and rice, rolled in cabbage or vine leaves, etc.'. Moreover, the hypernym of {сарма} (sarma), the synset {ястие; блюдо; гозба; манджа}(yastie, blyudo, gozba; mandzha) 'dish' defined as 'food prepared in a certain way by thermal or other type of cooking',

is linked with its translation equivalent in English and all these relations outline the position of a corresponding empty node in the English WordNet (Fig. 1).

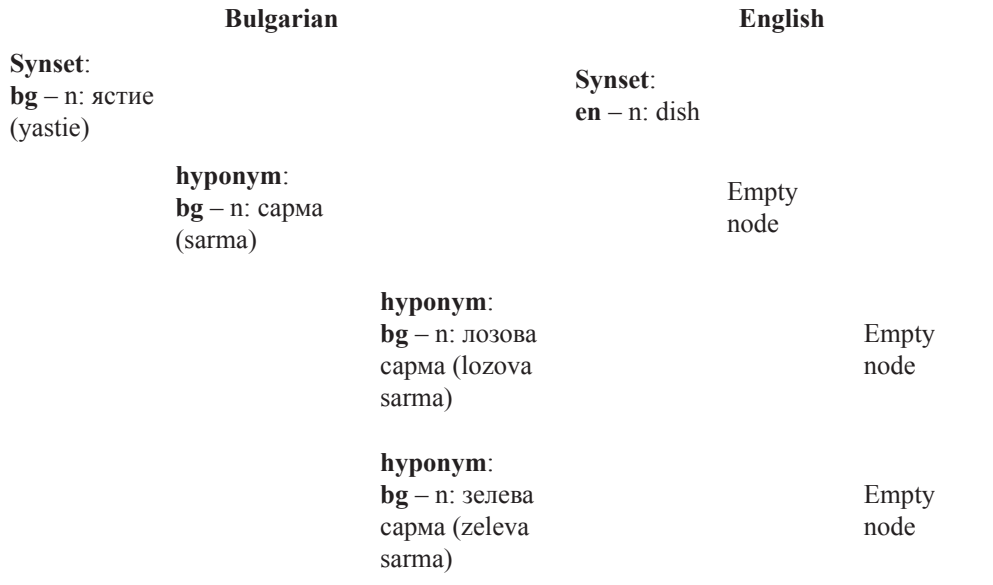

Fig. 1. Language specific lexicalization

Several thematic domains can be indicated as repleted with language-specific concepts: administrative and political system; kinship; religion and beliefs; architecture; folk music; animals and plants; food and beverage; clothing; traditional arts and crafts; work tools; units of measurement; holidays and significant historical events among others.

After Princeton WordNet, many (lexical-)semantic networks were developed for many languages, and this became a prerequisite for the creation of a Global WordNet (Multilingual Wordnet) (Bond et al. 2016), which is built around a shared set of "Base" Concepts. The notion of Base Concepts targets maximum overlap and compatibility across wordnets of different languages, allowing at the same time the distributive development of semantic networks with language-specific structure and lexicalization patterns (Vossen et al. 1998, 53). Ideally, the Inter-Lingual Index of the Global WordNet will include all concepts for which there is lexicalization in at least one language. 1,024 Base Concepts are identified initially on the basis of English, Dutch, Spanish and Italian along the following criteria: high position in the semantic hierarchy and maximum number of relations with other concepts in the network. New Base Concepts have been added of second and third order on the basis of data from Bulgarian, Greek, Romanian, Serbian, and Turkish (Tufiş, Cristea, Stamou 2004, 18) and the batch of Base Concepts has been expanded to 4,689, observing the following additional criteria: the most frequent words in large representative corpora and hypernyms of the already chosen synsets running up to the root of the trees (Koeva 2014, 160). The Base Concepts are part of the structure of the Bulgarian WordNet, which makes it part of the Global WordNet.

A Collaborative Inter-Lingual Index is developed that integrates the data for a single Inter-Lingual Index of concepts for wordnets such as the concepts should represent salient and frequent lexicalized concepts in all languages and the relations should have the same meaning for all languages (Bond et al. 2016). It is clear that some cross-language links might not express total equivalence between the meaning of the synsets in different languages in a similar way as the synonyms in a synset in most cases are cognitive or contextual synonyms and thus interchangeable in a finite number of contexts.

4. SynSet structure in the Bulgarian WordNet

The following definition for synonymy is applied to WordNet: "Two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value" (Miller et al. 1990/1993, 6). This test shows that the WordNet includes cognitive (or propositional) synonyms and (rarely) absolute synonyms. The absolute synonyms in WordNet are usually limited to spelling alternations and abbreviations (actually, the absolute synonyms may exist for a narrow timespan or for a narrow group of people). For example, in Bulgarian such are the literals {халище, кименик, губер} (halishte; kitenik, guber) 'shag, rug' with the definition 'a thick fluffy rug that serves as a mat or duvet'. Cognitive synonyms are words (varying in their evaluative or connotative meaning) that can be interchangeable in each grammatical declarative sentence (Cruse 1986, 86), while contextual synonyms are limited to a specific context (Lyons 1981, 149). For example, the pairs {assumption; supposition}, {race; run}, {delightful; delicious} are cognitive synonyms while the synonyms {educator, pedagogue, pedagog} are not interchangeable in many contexts: in the collocation with the word *certified* most preferred is the word *educator* (*certified educator*), followed by *pedagogue*, while the word *pedagog* is most rarely used; in the expression *Microsoft certified educator* the word *educator* could not be replaced with either of the words *pedagogue* or *pedagog*.

4.1. Types of literals

The literal can be a simple word, derived word or a compositional word (multiword expression), an abbreviation {**DOS**; disk operating system}, a symbol (oxygen; **O**), a formula {carbon dioxide; $\mathbf{CO_2}$}, or a digit {two; **2**; **II**}.

Multiword expressions (words denoting a single concept, but constituted by two or more orthographic words) can be identified on the basis of a number of tests: whether they have as a synonym a simple word: *правя избор* (pravya izbor) 'make a choice' and *избирам* (izbiram) 'choose'; whether they have as a translation equivalent a simple word in another language: *женско дружество* (zhensko druzhestvo) 'sorority' and *sorority*; whether they have as a synonym an abbreviation: *операционна система* (operacionna sistema) 'operating system' and *ОС* (OS) 'OS'; whether they represent a term or a proper noun: *европейски хамелеон* (evropeyski hameleon) 'European chameleon' and *Chamaeleo chamaeleon*.

There are numerous classifications of multiword expressions (Sag et al. 2002; Baldwin, Kim 2010; etc.). We present here a short classification with the view of illustrating the types of multiword expressions in the Bulgarian WordNet: a) expressions, which are semantically and syntactically compositional, but are used to denote a single concept, and b) expressions, which are fully lexicalized with different degrees of semantic and syntactic compositionality (non-transparency). The first type of multiword expressions is difficult to determine, their constituents express the meaning with which they participate both in the particular multiword expression and separately, but as a multiword expression they denote compositionally the meaning of a single concept and have been conventionalized for its naming: *пощенска кутия* (poshtenska kutiya) 'mail box', *консервна кутия* (konservna kutiya) 'can'. The lexicalized expressions on their part are: fixed – fully lexicalized expressions which do not change either in terms of word order or grammar: *благодарение на* (blagodarenie na) 'thanks to'; semi-fixed – the order and constituency of the parts of the compositional word are fixed but the constituents can undergo certain paradigmatic changes within certain grammatical categories: *полярна мечка* (polyarna mechka) 'polar bear', *полярни мечки* (polyarni mechki) 'polar bears'; *Ден на майката* (Den na maykata) 'Mother's day', *Денят на майката* (Denyat na maykata) 'the Mother's day'; non-fixed – they change morphologically, can undergo word order changes, as well as tolerate mutable elements in their composition: *изнасям лекция* (iznasyam lekciya) 'deliver a lecture'; *лекция ще изнасям утре* (lekciya shte iznasyam utre) 'will deliver a lecture tomorrow'; *изнасям доклад* (iznasyam doklad) 'deliver a paper'; *чета доклад* (cheta doklad) 'deliver a paper'. In the Bulgarian WordNet, the base form of the multiword expressions is recorded and additionally the head word and dependent parts are marked.

4.2. Lexical and stylistic labels

When there is a specific use – a substandard word, figurative meaning, phraseological meaning, an obsolete word, informal use, abbreviation and so on, a label is attached to the respective literal within a literal note. The system of labels developed for the Bulgarian WordNet reflects the following **differences in usage** of synonyms: **belonging to non-standard lexis** – a dialectal word: *кърчаг* (karchag) 'jug' instead of *стомна* (stomna) 'jug'; slang: *копеле* (kopele) 'bastard' instead of *незаконно дете* (nezakonno dete, literal translation – illegitimate child) 'bastard'; folk word: *бабина душица* (babina dushitsa, literal translation – grandma's soul) 'thyme' instead of *мащерка* (mashterka) 'thyme'; a word with an undesired use: *дрескод* (dreskod) 'dress code' instead of *облекло* (obleklo) 'dress code'; **use in a specific functional style** – a colloquial word: *дрънкам* (drankam) 'tweedle' for 'play negligently on a musical instrument'; a poetic word: *дихание* (dihanie) 'breath' instead of *дъх* (dah) 'breath'; a literary word: *мечтание* (mechtanie) 'revery' instead of *мечта* (mechta) 'dream'; term: *абдомен* (abdomen) 'abdomen' instead of *корем* (korem) 'stomach'; **the historical period of use** – an obsolete word: *крушение* (krushenie) 'wreck'; a historical word: *васал* (vasal) 'vassal'; a new word: *дрон* (dron) 'drone'; **the expressive properties of the literals** – a word with pejorative meaning: *писарушка* (pisarushka) 'pen-driver';

a collective word: *селячество* (selyachestvo) 'peasantry'; a diminutive word: *възглавничка* (vazglavnichka) 'cushion'[5]; **the frequency of use of the literals** – a rare word: всеядец (vseyadets) 'omnivore'; **the nuances in the use of the literals** – a figuratively used word: *връзвам*, *вържа* (vrazvam, varzha) 'tie' with the meaning 'limit or restrict to'; a doublet: *шейсет* (sheyset) 'sixty' and *шестдесет* (shestdeset) 'sixty'.

The **grammatical labels** indicate a constant grammatical feature of a particular literal. Perfective and imperfective verbs in Bulgarian express different meanings, although the verb aspect pairs are closely related: *родя* (rodya) 'give birth' and *раждам* (razhdam) 'am giving birth'. The different meaning of verb aspect pairs is reflected at both the morphological and the syntactic levels: the paradigms of the perfective and imperfective verbs are different – perfective verbs do not have the so-called independent present tense, they do not form present participles (agentive and adverbial), nor negative imperative forms; the derivational potential of the perfective and imperfective verbs is different – perfective verbs do not form some types of deverbal nouns and some nouns denoting professions; perfectivity is directly related to the syntactic realization of obligatory complements – direct objects of perfective verbs cannot remain implicit and perfect verbs cannot be complements of phase predicates; perfectivity is also directly related to the possibility for different types of verb diathesis: perfective verbs do not form middles, optatives, impersonals. In the Bulgarian WordNet verb aspect pairs are included in one and the same synsets, although the perfective and imperfective members of a pair are not cognitive synonymous, and as a consequence only some of the literals are translation equivalents to the respective synonyms in English. Such a representation is fully synchronous with the tradition in dictionary compilation of imperfective and perfective verbs being included in a single dictionary entry. However, the main motivation is preserving the structure of the Princeton WordNet. For the differentiation of verbs of different aspect, a literal note is attached to each verb indicating its aspect: perfective verb: *запея* (zapeya) 'start singing'; imperfective verb: *запявам* (zapyavam) 'sing off'; a simultaneously perfective and imperfective verb: *пенсионирам* (pensioniram) 'retire'; an imperfective verb with no perfective equivalent: *вали* (vali) 'rain'; a perfective verb with no imperfective equivalent: *повярвам* (povyarvam) 'get to believe'. The literal notes allow further splitting of the Bulgarian verb synsets according to different verb aspects, linking one of the resulting synsets with the English synset equivalent and introducing a new aspectual relation between the split synsets.

The grammatical labels also describe some paradigmatic restrictions: a word used exclusively in the singular: *юдаизъм* (yudaizam) 'Judaism'; a word typically used in the singular: *критика* (kritika) 'criticism'; a word exclusively used in the plural: *финанси* (finansi) 'finances'; a word typically used in the

[5] The formation of diminutive words in Bulgarian is regular, in Bulgarian WordNet the diminutive words are included if they have a new meaning built on the diminutive one.

plural: *зелева сарма* (zeleva sarma) 'cabbage roll'; a word exclusively used with a definite article: *Острова* (Ostrova, literal translation – the island) 'the United Kingdom'; a word typically used with a definite article: *Десетте Божи заповеди* (Desette Bozhi zapowedi) 'the Ten Commandments'.

The grammatical labels provide also information for the part of speech of the literals, if it is different from the part of speech of the synset in English: substantivized adjective or participle: *свинско* (svinsko) 'pork', an adjective in Bulgarian for *свинско месо* (svinsko meso) 'pork'; pronoun: *някакъв* (nyakakav) 'some', a pronoun in Bulgarian for *some*, qualified as an adjective in the Princeton WordNet with a definition 'quantifier; used with either mass nouns or plural count nouns to indicate an unspecified number or quantity'; a prepositional phrase: *с усмивка* (s usmivka) 'with a smile', a prepositional phrase in Bulgarian which corresponds to the English adverb *smilingly* with a definition 'with smiles; in a smiling manner'.

Also some labels indicate **semantic and syntactic compatibility between words**: in combination with a preposition: *запознат* (zapoznat) 'acquainted' is used with the preposition *с* (s) 'with'; in combination with an adverb: *разраствам се* (razrastvam se) 'proliferate' is used with the adverb *бързо* (barzo) 'rapidly'; in combination with a noun: *драсвам* (drasvam) 'write, to drop a line' is used with the noun *бележка* (belezhka) 'message'.

The labels describing the use of literals specify the context they can be encountered in. The **graphemic difference** is also marked in the notes: a contraction: *гр.* (gr.) 'city' instead of *град* (grad) 'city'; an abbreviation: *БНБ* (BNB) 'Bulgarian National Bank' instead of *Българска народна банка* (Balgarska narodna banka) 'Bulgarian National Bank'; Latin spelling: *Pogostemon cablin* for 'small East Indian shrubby mint; fragrant oil from its leaves is used in perfumes'. The multiword expressions: *сключвам мир* (sklyuchvam mir) 'make peace', are also marked with the label phraseological expression.

4.3. Definition of synsets

Unlike some of the definitions in the Princeton WordNet and in some other wordnets, where the convention has been adopted to present a short definition which indicates but does not fully describe the meaning, the definitions in the Bulgarian WordNet are characterized by: a) representation of the meaning in the manner adopted for the definitions in monolingual general dictionaries; b) originality of the definition, i.e., lack of repetition of the definitions provided in existing dictionaries or encyclopaedias (this is achieved in many cases through the expansion of the definitions by the addition of further information); c) the observation of a predetermined structure, which is applied not only to different parts of speech, but also to synsets belonging to a given thematic domain.

As has been customarily established, the definition of nouns has a generalizing part, which contains the hypernym (in some cases not the immediate one), and a specifying part, which renders the meaning unique. The definition of verbs is subject to the same structure and, when possible, contains information on the arguments. In taxonomic classifications, it is possible for a holonym to

be used in the generalizing part (for example, *a representative of the class Mammals*). The definition with adjectives starts with *който е* (koyto e) 'who is', *който се характеризира с* (koyto se harakterizira s) 'who is characterized by'. The definition of relational adjectives contains the noun from which the adjective has been derived and the standard definition includes expressions such as: *който се отнася до* (koyto se otnasya do) 'who refers to'; *който се свързва с* (koyto se svarzva s) 'who is related to'; *който е характерен за* (koyto e harakteren za) 'who is characteristic of' and other similar definitional descriptions. The definition of adverbs starts with: *означава, че* (oznachava, che) 'means that'; *характеризира ... като* (harakterizira ... kato) 'characterizes ... as'. The definitions do not contain any of the synonymous literals; therefore there are no definitions with synonyms.

A common general structure of the definition is followed in the description of the synsets in a given thematic domain, where the individual definitions may not contain all the elements from the common structure. For example, the definitions for plants include the following elements: membership to a given taxonomic unit, geographic distribution, physical characteristics, economic significance, e.g., the synset {райска ябълка} (rayska yabalka) 'oriental persimmon' is described by the definition 'a representative of the tree plants of the family Ebenaceae with edible, sweet, yellow to orange fruit'.

With terms, the thematic domain is indicated only if all the literals in a synset are terms, for example, for the synset {шипка; шипков храст; Rosa eglanteria; Rosa rubiginosa} (shipka; shipkov hrast) 'brier; brier bush' the definition starts with the expression *в ботаниката* – (v botanikata –) 'in botanics –'. If there are further restrictions in use, these are explicated before the definition, for example, *в анатомията, при някои животни* – (v anatomiyata, pri nyakoi zhivotni –) 'in anatomy, for certain animals –'. One hundred and thirty-six thematic domains have been systematized in the Bulgarian WordNet within the four spheres of human activity: science, art, religion, and sports.

4.4. Usage examples of synonyms

In the creation of synsets, at least one usage example is added, which showcases the meaning of the synsets. The examples are in the form of (in the usual case) full sentences and created in a few specific ways: as a translation of the example in the Princeton WordNet, via the search and choice from corpora, or through dedicated construction. The multitude of examples in the different languages can be thought of as a multilingual corpus with corresponding sentences with semantic annotations for the respective literals.

4.5. Organization of the synsets in semantic classes

The nouns in the Princeton WordNet are organized into semantic classes (primitives): generic concepts, perceived as unique roots (beginners) of separate hierarchies, and the nouns belonging to the hierarchies are subsumed under the common semantic class (Miller 1990/1993, 16). Twenty-five semantic classes have been chosen, as follows: {plant, flora}, {animal, fauna}, {person, human

being}, {body, corpus}, {natural object}, {artifact}, {natural phenomenon}, {substance}, {food}, {possession}, {process}, {quantity, amount}, {relation}, {shape}, {state, condition}, {act, action, activity}, {attribute, property}, {cognition, knowledge}, {communication}, {event, happening}, {feeling, emotion}, {group, collection}, {time}, {location, place}, {motive}. The verbs are also organized into semantic classes, numbering fifteen such altogether: {bodily care and functions}, {change}, {cognition}, {communication}, {competition}, {consumption}, {contact}, {creation}, {emotion}, {motion}, {perception}, {possession}, {social interaction}, {weather verbs}, {state} (Fellbaum 1990/1993, 47).

The semantic classes of nouns and verbs may be specified into a number of semantic subclasses. For example, within the semantic class {храна} (hrana) 'food' we can introduce a subclass {напитка} (napitka) 'drink' for nouns that can be associated with verbs such as *отпивам* (otpivam) 'have a sip', *пия* (piya) 'drink' and so on. The semantic classes of nouns in the Bulgarian WordNet are organized in a hierarchy of 253 semantic subclasses via the establishment of correspondences with the ontology of the semantic types in the Corpus Pattern Analysis (Hanks 2012, 57-59). The semantic types refer to characteristics which the nouns, combining with specific verbs, should possess. The semantic types are part of verb patterns – abstract structures which specify the semantic preferences of the verbs in relation to synsets of nouns and noun phrases:

([Human] | [Animal]) loves ([Activity] | infinitival clause, for Bulgarian a *da*-clause)

For example, *Момчето обича играта* (Momcheto obicha igrata) 'The boy loves the play', *Котката обича да играе* (Kotkata obicha da igrae) 'The cat loves to play'.

The semantic types have been manually mapped with the appropriate synsets as follows: 199 semantic types are mapped with one synset: for example, the type [Permission] to the synset {permission} 'approval to do something' with the semantic class {communication}; the type [Dispute] to the synset {disagreement} 'the speech act of disagreeing or arguing or disputing' with the semantic class {communication}; 39 semantic types are correlated to two synsets: for example, the type [Route] to the synset {road; route} 'an open way (generally public) for travel or transportation' with the semantic class {artefact} and to the synset {path; route; itinerary} 'an established line of travel or access' with the semantic class {location}; 12 semantic types are correlated to three synsets; 2 semantic types are correlated to four synsets and 1 semantic type – to five synsets (Koeva et al. 2018). An automatic mapping procedure has been applied for correlating the ascribed semantic types to the corresponding hyponym synsets, which results in 172,991 new mappings[6]. The more detailed presentation of the semantic classes in the ontological structure allows the inheritance between semantic classes down the hierarchy and guarantees more precise predictions of the combinability between a verb and nouns in language expressions.

[6] The mapping is available at: https://dcl.bas.bg/PWN_CPA/

The adjectives in the Princeton WordNet are divided into three semantic classes: relational adjectives, which have been derived from nouns (most of these are relational adjectives in Bulgarian); descriptive adjectives (all remaining adjectives) and a small group of participles with adjectival functions. A classification has been developed (Dimitrova, Stefanova 2018), based on the semantic classification used in the German WordNet, which is motivated by the fact that the modifying adjective is (semantically) related in a specific way with a given (modified) noun (Hundsnurscher, Splett 1982). The classification for adjectives introduced in the Bulgarian WordNet has seventeen semantic classes, as follows: **social** or **public membership** – {законен} (zakonen) 'legal', {пролетарски} (proletarski) 'proletarian'; **place** or **location** – {съседен} (saseden) 'neighbouring', {централен} (tsentralen) 'central'; (astronomical) **time** – {едночасов} (ednochasov) 'one-hour', {дългосрочен} (dalgosrochen) 'long-term'; **atmospheric conditions** or **climate** – {мъглив} (magliv) 'foggy', {дъждовен} (dazhdoven) 'rainy'; **physical characteristic** – {мускулест} (muskulest) 'muscular', {широколистен} (shirokolisten) 'broadleaf'; **sensory perception** – {горещ} (goresht) 'hot', {шумен} (shumen) 'noise'; **movement** – {подвижен} (podvizhen) 'mobile', {тичащ} (tichasht) 'running'; **quantity** – {многоброен} (mnogobroen) 'numerous', {двоен} (dvoen) 'double'; **knowledge** or **cognition** – {логически} (logicheski) 'logical', {неокласически} (neoklasicheski) 'neoclassical'; **relation** – {подотраслов} (podotraslov) 'sub-branch', {подготвителен} (podgotvitelen) 'preparatory'; **feeling** or **emotion** – {уплашен} (uplashen) 'scared', {бесен} (besen) 'furious'; **behaviour** – {дипломатичен} (diplomatichen) 'diplomatic', {послушен} (poslushen) 'docile, obedient'; **substance** or **compound** – {алкален} (alkalen) 'alkine', {йодиращ} (yodirasht) 'iodising'; **material** – {кожен} (kozhen) 'leather', {железен} (zhelezen) 'iron'. There are separate semantic classes relating to the qualitative characteristics of animate and inanimate entities: **quality** – {остър} (ostar) 'sharp', {противен} (protiven) 'obnoxious'; to a **change of state** – {вцепеняващ} (vtsepenyavasht) 'benumbing', {сковаващ} (skovavasht) 'paralysing'; as well as to a particular **state** – {спящ} (spyasht) 'sleeping', {счупен} (schupen) 'broken'. In the Bulgarian WordNet 4,182 adjectival synsets have been classified within some of the above listed semantic classes.

The adverbs in the Bulgarian WordNet are subdivided in 9 semantic classes depending on their most general categorical meaning (time, place, spatial orientation, manner, degree, frequency, quantity, modality, logical relation). Each adjective synset is linked to the noun synset denoting the respective **thematic domain** (for example, the synset {ежедневно; всекидневно} (ezhednevno; vsekidnevno) 'daily' is linked to the synset {честота, честотност, фреквентност} (chestota; chestotnost; frekventnost) 'frequency; frequence; oftenness' and to the synset {време} (vreme) 'time' with the relations **domain category**. Thus, adverbs in the Bulgarian WordNet are practically organized into semantic classes (similarly to nouns, verbs and adjectives). Some adverbs are also associated with the extralinguistic relation **domain usage** in accordance with their use.

4.6. Sentence frames of verb synsets

The Sentence frames in the WordNet have the following format: *Somebody ----s something to somebody* (Fellbaum 1990/1993: 45), in which, beside the number and type of the syntactically obligatory constituents, also included is minimal information about the selective restrictions on combinability: whether a given element in the frame can be realized as a noun that is human or not, and minimal syntactic information: whether the element is realized as a noun, prepositional phrase or a subordinate clause. For example, the synset {hate; detest} with a definition 'dislike intensely; feel antipathy or aversion towards' is associated with the Sentence frame: *Somebody ----s somebody* and *Somebody ----s something*. In the Princeton WordNet 35 generic frames have been provided and they are ascribed either to the synset (which means that the respective frame applies to all verbs, members of the synset) or only to some of the verbs in the synset, if there are differences in combinability. The Semantic frames of FrameNet (Ruppenhofer et al. 2016) have been automatically mapped to the Bulgarian WordNet (the Semantic frames contain detailed semantic and syntactic information about the elements in the frame). The WordNet to FrameNet mapping is based on three previous lexical mappings: 2,817 direct mappings provided within FrameNet (Baker, Fellbaum 2009), 3,134 from eXtendedWordFrameNet (Laparra, Rigau 2010), and 1,833 from MapNet (Tonelli, Pighin 2009), and on 1,335 structural mappings with VerbNet (Shi, Mihalcea 2005). All in all, the unification of mappings resulted in 4,306 unique mappings of a WordNet synset onto a FrameNet Semantic frame (Leseva, Stoyanova 2020). As a general procedure, the hypernym's frame was transferred to its hyponyms in the cases where the hyponyms are not directly mapped to FrameNet frames. As a result, 13,226 synsets were automatically assigned a FrameNet Semantic frame. After the automated mapping, a series of procedures have been carried out for improving the results: searching for an additional match between literals in the given synset and the FrameNet lexical units in the related and sister frames; calculation of similarity between the gloss of a verb synset and FrameNet lexical unit definitions, searching for a match between literals and words contained in the FrameNet frame name. As a result of these steps, 9,341 new suggestions for more specific or other possible frames have been made for 5,661 synsets. Among all mappings 5,025 frames assigned to verb synsets in WordNet have been manually validated by experts[7].

5. Relations in WordNet

It was stated that two words W and W_2, denoting respectively sets of entities E and E_2, can be in one of the following relations (Cruse 1986, 87): identity: $E = E_2$; inclusion: E_2 is contained in E; overlapping: E and E_2 have a non-empty intersection, but are not included into one another; exclusion: E and E_2 share no element. Respectively, **homonyms** are in a mutually exclusive relation, **poly-**

[7] The mapping is available at: https://dcl.bas.bg/en/semantic-relations-data/

semous words are in a relation of inclusion or overlap; **synonyms** – in relation of identity; **antonyms**, relations **verb groups**, **also see** and **similar to** – in relation of overlap; **hypernyms**, **meronyms** and **subevents** – in relation of inclusion. Depending on their properties, the relations project or not hierarchical structures.

Hierarchical relations (relations of inclusion) are of three basic types – taxonomic (classificatory, which associate an entity of a particular type with an entity of a more generic type), meronymic (expressing the relation of the whole to its parts) and proportional series (expressing proportions between values in a given series) (Cruse 1986, 112-145). Non-hierarchical relations are binary or tertiary (for example, **synonymy** can be viewed as a tertiary relation – word W and word W_2 are synonyms in context C, even though, for simplicity, it is interpreted as a binary relation).

Taxonomic relations are: inverse and transitive (**is hypernym of** and **has hypernym**, **has hyponym** and **is hyponym of; has troponym** and **is troponym of**); meronomic relations are also inverse and transitive (**is holonym of** and **has holonym**, **has meronym** and **is meronym of**, **has subevent** and **is subevent of**); and inverse and non-transitive (extralinguistic, morpho-semantic, and causative relations). Non-hierarchical relations are: symmetric, reflexive, transitive and Euclidean (**synonymy**), symmetric, irreflexive and non-transitive (**antonymy**); symmetric, irreflexive and Euclidean (**similar to**, **also see**, **verb group**), and inverse and transitive (derivational relations).

For example, the synset {cookbook, cookery book} with the definition 'a book of recipes and cooking directions' is a hyponym of the synset {reference book} with the definition 'the contributed articles to the basic reference work on that topic', which in turn is a hypernym of the former, as the relation is inverse, and a hyponym of {book} 'a written work or composition that has been published (printed on pages bound together)'. The transitivity of the relation is illustrated by the fact that {book} is a hypernym of both {reference book} and {cook book}. The hypernym can replace the hyponym in some context, but not vice versa: *The cookbook* (*reference book, book*) *is useful for any housewife who wants to please her husband with a delicious moussaka*; *The book* (*the cookbook) *tells of the adventures of two tiger cubs*.

Several types of **meronymy** can be identified (Koeva, Stefanova 2020), three of which are defined in the WordNet: **part of**, **member of**, and **substance of**. The relation **part of** connects parts to their whole only if there is a topological inclusion through physical contact: *a book is part of a library*, *a library is part of a building*, **a book is part of a building*. For the relation **member of** there is a condition that the holonym should be a collective noun or a noun denoting a set of objects: *basketball player – basketball team – basketball league*. For the relation **substance of**, the condition is that the holonym must express a material or a substance while the meronyms express the ingredients. The components might not be expressed by a word, but descriptively, through a phrase that refers to the whole: *cake – a piece of cake*.

The batch of relations in the Princeton WordNet has been preserved in the Bulgarian one, with a few exceptions: some relations have been renamed, for example **substance homonymy** / **meronymy** is labelled **portion homonymy** /

meronymy, **troponymy** in verbs is labelled **hyponymy**, while **entailment** (proper inclusion) is defined as **subevent** and **entailment** (backward presupposition) has not been transferred (Piasecki, Koeva 2017); **antonymy** and derivational relations in English are included in the Bulgarian WordNet at the level of the synsets and **antonymy** can be characterized as **near antonymy**. Derivational relations transferred from the Princeton WordNet are encoded as symmetric relations since the direction of the derivation is not indicated. The addition of new types of relations is restricted at present to the addition of derivational and morpho-semantic (syntagmatic) relations. The relations in the Bulgarian WordNet in March 2021 are 254,821, among which 246,517 relations between synsets and 8,304 relations between literals (not taking into account the **synonymy** relation). The lexical, conceptual and extralinguistic relations in the Bulgarian WordNet are presented in Table 2.

Table 2. Description of the relations in the Bulgarian WordNet

Relation	Meaning	Examples
Lexical relations		
Semantic relations		
Synonymy (**is synonym of**)	Expresses equivalence of meaning between pairs of nouns, of verbs, of adjectives or of adverbs. In the Bulgarian WordNet synonymous sets of pronouns, prepositions, conjunctions, particles and interjections are introduced.	*newspaper – paper*; *learn – study*; *nice – beautiful*; *quickly – rapidly*; *някакъв* (nyakakav) 'any'– *някой* (nyakoy) 'any'; *върху* (varhu) 'on' – *на* (na) 'on'; *и* (i) 'and' – *та* (ta) 'and'; *ли* (li) 'whether' – *дали* (dali) 'whether'; *о* (o) 'ah' – *ох* (oh) 'ah'
Antonymy (**is antonym of**)	Expresses oppositeness of meaning between pairs of nouns, of verbs, of adjectives or of adverbs.	*day – night*; *buy – sell*; *handsome – ugly*; *quickly – slowly*
Derivational relations		
Participle (**has participle – is participle of**)	Expresses the derivation of participles (functioning as adjectives) from verbs.	*saponify – saponified; thoriated – thoriate*

Table 2 (continued)

Relation	Meaning	Examples
Derivative (**has derivative – is derivative of**)	Expresses the derivation of nouns from verbs.	*wait – waiter*
Pertainym **has pertainym – is pertainym of**)	Expresses the derivation of adjectives from nouns.	*atom – atomic; academic – academia*
Conceptual relations		
Semantic relations		
Hypernymy (**is hypernym of – has hypernym**) Hyponymy **(is hyponym of – has hyponym**)	Expresses the inclusion of classes (genus-kind relations) for nouns.	*animal – elephant; school bell – bell*
Hypernymy (**is hypernym of – has hypernym**) Troponymy (**is troponym of – has troponym**)	Expresses the relation between a more general and a more specific manner of doing something for verbs. In the Bulgarian WordNet the term Hyponymy is used for Troponymy.	*love – adore; limp – walk*
Instant hypernymy (**is instant hypernym of – has instant hypernym**) Instant hyponymy (**is instant hyponym of – has instant hyponym**)	Expresses the relation between a noun describing the type of a proper noun and the proper noun.	*a Balkan country – Bulgaria*; *Shakespeare – dramatist*
Holonymy (**is holonym of – has holonym**) (Meronymy) (**has meronym – is meronym of**)	Expresses a relation between a whole and its parts (encoded by nouns). There are **part**, **member** and **substance** holonymy and meronymy. In the Bulgarian WordNet the term portion is used to replace the term substance.	*building – room; eyelid – eyelash; football team – football league; army corps – division; pastry – flour; flour – bread*
Subevent (**has subevent – is subevent of**)	Expresses temporal inclusion in verbs.	*buy – pay*; *mature – senesce*
Cause (**causes – is caused by**)	Expresses a cause-effect relation in verbs.	*blow up – explode*; *break* 'go to pieces' – *break* 'ruin completely'

Table 2 (continued)

Relation	Meaning	Examples
Attribute (**has attribute – is value of**)	Expresses values of a noun through adjectives.	*weight – light, heavy*; *high, low – height*
Similar to (**is similar to**)	Expresses similarity in the meanings of a "central" adjectival synset and a satellite adjectival synset.	*savage – wild*
Verb group (**is in a verb group with**)	Expresses a relationship of similarity between verb synsets.	*act – behave*
Also see (**is also see of**)	Expresses the association of meanings between adjectival synsets.	*urgent – imperative*
Extralinguistic relations		
Domain category (**is a domain category of – has a domain category**)	Expresses the relation between a synset (noun) denoting a specialized thematic area and a synset (noun, verb, adjective or adverb) which is used in the respective thematic area.	*law – diplomatic immunity; biology – inhibit; medicine – epidemic; biology – symbiotically*
Domain region (**is a domain region of – has a domain region**)	Expresses the belonging of the synset (noun) to a particular geographic or cultural-geographic region (noun).	*France – French Revolution; steppe – Russia*
Domain usage (**is a domain usage of – has a domain usage**)	Expresses the relation between a synset (noun) for a specialized sphere of use and a synset (noun, verb, adjective and adverb) with a specific use in the respective sphere.	*slang – hood* 'a neighborhood'; *slang – chuck* 'throw away'; *slang – grotty* 'very unpleasant or offensive'; *slang – clean* 'completely; used as intensifiers'

The derivational relations that have been transferred from the Princeton WordNet are at the level of the synset and the most frequent general relation (named **derivationally related form**) is labelled as **derivative in English**. Part of the derivational relations have been checked and if the relation has proven valid for Bulgarian, the relation **derivative in Bulgarian** has been introduced. For example, between the noun {тостер} (toster) 'toaster' with the meaning 'a kitchen appliance (usually electric) for toasting bread' and the literals from the synset {препичам; препека} (prepicham; prepeka) 'toast' with the meaning 'for bread – make brown and crisp by heating' **there is no derivational relation, unlike** in English, but between the noun {препичане} (prepichane) 'toasting' and the literal *препичам* (prepicham) 'toast' form the synset {препичам; препека} (prepicham; prepeka) 'toast' in Bulgarian, there is a derivational relation.

Since derivational relations are important for determining morpho-semantic relations, selected derivational relations were introduced at the level of literals in the Bulgarian WordNet. These derivational relations do not show the semantics of the derivation, rather they reflect the word-formation mechanism: suffixation: *плувам* (pluvam) 'swim' – *плуване* (pluvame) 'swimming'; conversion: *вечерям* (vecheryam) 'dine' – *вечеря* (vecherya) 'dinner'; prefixation: *винт* (vint) 'screw, bolt' – *завинтя* (zavintya) 'screw in'; substitution (symmetrical): (verb suffix – nominal suffix): *акомпанирам* (akompaniram) 'accompany' – *акомпанимент* (akompaniment) 'accompaniment'; derivation: (a non-specified derivational relation, which is not immediately obvious from a contemporary point of view or is not part of the Bulgarian system of word formation): *помогна* (pomogna) 'help – a verb' – *помощ* (pomosht) 'help'; *мириша* (mirisha) 'smell – a verb' – *миризма* (mirizma) 'smell'. On the other hand, the semantics of the derivation is expressed by the derivational means: different suffixes and prefixes, although they are ambiguous. The derivational relations between a verb and a noun have been automatically predicted and manually checked as a result of which there are 6,220 new derivational relations ascribed in the Bulgarian WordNet (Dimitrova, Tarpomanova, Rizov 2014).

If there exists a derivational relation between two synsets connecting some of the literals that make them up, then there exists a semantic relation of a given type between these two synsets (called a morpho-semantic relation), for example: **agent** (*a teacher teaches*; *a decorator decorates*; *a leader leads*), **instrument** (*a protector protects*; *a chisel chisels*), and so on. These relations are grounded in the idea that meanings of word-formation affixes can be exhaustively classified in a small number of semantic categories of the type 'semantic role' (Fellbaum, Osherson, Clark 2009). The morpho-semantic relations between two synsets, unlike derivational relations, apply to all literals constituting the respective synsets (*a protector protects* but also *a fuse protects*). What is more, morpho-semantic relations are language independent, that is if such relations are identified with the help of language-specific derivation, this may be used as the basis for their transfer in other languages, where lexicalizing mechanisms are different.

A morpho-semantic relation can be expressed by different derivation mechanisms. For example, the word *певец* (pevets) 'singer' from the synonymous set {певец, вокалист} (pevets; vokalist) 'singer, vocalist' with the definition 'musician who performs vocal parts in musical works' is formed with the suffix *-ец* (-ets) from the literal {*пея*} (peya) 'sing' with the meaning 'create melodic tones with his voice', while the second literal *вокалист* (vokalist) 'vocalist' is formed with the suffix *-ист* (-ist) from the literal {вокализирам} (vokaliziram) 'vocalize' with the definition 'sing (each note on a scale or in a melody) with the same vowel'. For English, the following semantic relations are defined for derivationally related verbs and nouns ending in *-er* / *-or*: **agent**: *dance – dancer*; **instrument**: *aspirate – aspirator*; **material**: *soften – softener*; **means**: *line – liner*; **vehicle**: *cruise – cruiser*; **location**: *plant – planter*; **undergoer**: *broil – broiler*; **body part actor**: *adduct – adductor* (for muscle). Semantic relations as a result of other word-formation mechanisms in English are: **event**: *dance – dance* (noun); **state**: *silence – silence* (noun); **result**: *inconvenience – inconvenience* (noun); **property**: *prevail – prevalence*; **uses**: *undercoat – undercoat* (noun); **means**: *warrant – warrantee*.

The defined morpho-semnatic relations in the Princeton WordNet were transferred to the Bulgarian one. However, a considerable number of derivationally related noun synsets to verb synsets both in the Princeton and the Bulgarian WordNet remained not labelled with a morpho-semantic relation. We developed a method for an automatic detection of derivationally related verb – noun pairs in the Bulgarian WordNet and for an automatic prediction and labelling of morpho-semantic relations between corresponding synsets (Koeva et al. 2016). Manual checking of the derivation relations in the Bulgarian WordNet confirmed the existence of the respective morpho-semantic relations.

6. Conclusion

The modelling of the information about the relations among concepts, among words, and between words and concepts in the Bulgarian WordNet (as in the other wordnets) has various applications in natural language processing: for the automatic extraction of the relations between different entities and automatic semantic role labelling; for the purposes of information and knowledge search and extraction; for automatic sense disambiguation; for the categorization of texts to different thematic domains; for text summarization; machine translation and various other uses.

The Bulgarian WordNet can be also used as an explanatory dictionary, a dictionary of synonyms or antonyms and as a bilingual or monolingual dictionary[8] within various use cases. A quite large part of the vocabulary in the Bulgarian WordNet has not been described so far in lexicographic resources or, if it has been described, the representation is not so complex or lacks a focus on semantic relations. Further, the Bulgarian WordNet has been mapped with the wordnets for 22 languages and became the biggest multilingual dictionary for Bulgarian: Albanian, Basque, Catalan, Croatian, Danish, Dutch, English, Finnish, French, Greek, Hebrew, Icelandic, Italian, Lithuanian, Polish, Portuguese, Romanian, Serbian, Slovak, Slovene, Spanish, Swedish[9] (and for some of these language pairs there are no available bilingual dictionaries to provide the translation units between the languages).

We can conclude that the applications of the WordNet are expanded much further than natural language processing by providing users a diverse lexicographic information in a systematic and easy to assess way. Comparing the Bulgarian WordNet to the Princeton WordNet reveals that the former was extended with the lexical and semantic information for lexical units representing all parts of speech; with labels pointing to the stylistic, lexical and grammatical features of words, with lexicographic definitions and examples, with detailed represen-

[8] https://dcl.bas.bg/bulnet/

[9] Data on the separate languages (to the exception of Bulgarian) are either distributed by the respective creators, holders of licences, which determine the way these can be used, or are used as the result of signed agreements with the copyright holders (for Romanian and Serbian).

tation of semantic classes of nouns, adjectives and adverbs, with the language specific derivational relations which indicate morpho-semantic relations, with points to the syntagmatic relations between sets of semantically and syntactically compatible verbs and nouns.

References

Baker, Fellbaum 2009: C. Baker, C. Fellbaum. WordNet and FrameNet as complementary resources for annotation. – Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP'09), ACL, Stroudsburg, PA, USA, 2009, 125-129.

Baldwin, Kim 2010: T. Baldwin. S. N. Kim. Multiword expressions. – In: Handbook of Natural Language Processing. Second Edition. Chapman and Hall/CRC, 2010, 267-292.

Bond et al. 2016: F. Bond, P. Vossen, J. McCrae, C. Fellbaum. CILI: The collaborative interlingual index. – Proceedings of the 8th Global Wordnet Conference (GWC 2016), 2016, 50-57.

Cruse 1986: D. A. Cruse. Lexical Semantics. Cambridge: Cambridge University Press, 1986.

Dimitrova, Stefanova 2018: T. Dimitrova, V. Stefanova. Semantic classification of adjectives in the Bulgarian Wordnet: Toward a multiclass approach. – Etudes Cognitives, 18, 2018, 1-17.

Dimitrova, Tarpomanova, Rizov 2014: T. Dimitrova, E. Tarpomanova, B. Rizov. Coping with derivation in the Bulgarian WordNet. – Proceedings of the Seventh Global Wordnet Conference, Tartu, Estonia, January 25-29, 2014, 109-117.

Fellbaum 1990/1993: C. Fellbaum. English verbs as a semantic net. – International Journal of Lexicography, 3, Winter 1990, 4, 278-301; reprinted in 1993, 40-51.

Fellbaum, Gross, Miller 1990/1993: C. Fellbaum, D. Gross, K. Miller. Adjectives in WordNet. – International Journal of Lexicography, 3, 1990, 4, 268-277; reprinted in 1993, 26-39.

Fellbaum, Osherson, Clark 2009: C. Fellbaum, A. Osherson, P. E. Clark. Putting semantics into WordNet's "Morphosemantic" Links. – In: Z. Vetulani, H. Uszkoreit (eds.). Human Language Technology. Challenges of the Information Society. LTC 2007. Lecture Notes in Computer Science, 5603. Berlin, Heidelberg: Springer, 2009, 350-358.

Hanks 2012: P. Hanks. How people use words to make meanings: Semantic types meet valencies. – In: J. Thomas, A. Boulton (eds.). Input, Process and Product: Developments in Teaching and Language Corpora. Brno: Masaryk University Press, 2012, 54-70.

Hundsnurscher, Splett 1982: F. Hundsnurscher, J. Splett. Semantik der Adjektive im Deutschen. Analyse der semantischen Relationen. Wiesbaden: Westdeutsches Verlag, 1982.

Koeva 2014: С. Коева. Wordnet и Булнет. – В: С. Коева (съст.). Езикови ресурси и технологии за българския език. София: Академично издателство „Проф. Марин Дринов", 2014, 154-173. (S. Koeva. WordNet i BulNet. – V: S. Koeva (sast.). Ezikovi resursi i tehnologii za balgarskiya ezik. Sofia: Akademichno izdatelstvo "Prof. Marin Drinov", 2014, 154-173).

Koeva, Stefanova 2020: С. Коева, В. Стефанова. Меронимията в УърдНет: дефиниране на субрелации. – В: В. Мичева, Д. Благоева, М. Витанова, М. Цибранска, С. Колковска, Т. Александрова (ред.). Доклади от Международната годишна конференция на Института за български език

„Проф. Любомир Андрейчин“ (София, 2020). Част 2. София: Издателство на БАН „Проф. „Марин Дринов“, 2020, 212-223. (S. Koeva, V. Stefanova. Meronimiyata v WordNet: definirane na subrelatsii. – V: V. Micheva, D. Blagoeva, M. Vitanova, M. Tsibranska, S. Kolkovska, T. Aleksandrova (red.). Dokladi ot Mezhdunarodnata godishna konferentsiya na Instituta za balgarski ezik “Prof. Lyubomir Andreychin” (Sofia, 2020). Chast 2. Sofia: Izdatelstvo na BAN “Prof. Marin Drinov”, 2020, 212-223.)

Koeva, Totkov, Genov 2004: S. Koeva, G. Totkov, A. Genov. Towards Bulgarian WordNet. – Romanian Journal of Information Science and Technology, D. Tufis, (ed.). Special Issue on BalkaNet, Romanian Academy, 7, 2004, 1-2, 45-61.

Koeva et al. 2016: S. Koeva, S. Leseva, I. Stoyanova, T. Dimitrova, M. Todorova. Automatic prediction of morphosemantic relations. – Proceedings of the Eighth Global WordNet Conference, Bucharest: RACAI, Global WordNet Association, 2016, 168-176.

Koeva et al. 2018: S. Koeva, T. Dimitrova, V. Stefanova, D. Hristov. Mapping WordNet concepts with CPA ontology. – Proceedings of the 9th Global WordNet Conference (GWC’2018). Global WordNet Association, Singapore, 2018, 70-77.

Laparra, Rigau 2010: E. Laparra, G. Rigau. eXtended WordFrameNet. – Proceedings of the Seventh International Conference on Language Resources and Evaluation, 17-3 May 2010, Valletta, Malta, 2010, 1214-1219.

Leseva, Stoyanova 2020: S. Leseva, I. Stoyanova. Beyond lexical and semantic resources: Linking WordNet with FrameNet and enhancing synsets with conceptual frames. – In: S. Koeva (ed.). Towards a Semantic Network Enriched with a Variety of Semantic Relations. Sofia: Professor Marin Drinov Publishing House of BAS, 2020, 21-48.

Lyons 1981: J. Lyons. Language and Linguistics. Cambridge: Cambridge University Press, 1981.

Miller 1986: G. Miller. Dictionaries in the mind. – Language and Cognitive Processes, 1986, 1, 171-185.

Miller 1990/1993: G. Miller. Nouns in WordNet: A lexical inheritance system. – International Journal of Lexicography, 3, 1990, 4, 245-264; reprinted in 1993, 10-25.

Miller et al. 1900/1993: G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller. Introduction to WordNet: An on-line lexical database. – International Journal of Lexicography, 3, 1990, 4, 235-244; reprinted in 1993, 1-9.

Piasecki, Koeva 2017: M. Piasecki, S. Koeva. WordNet Relations in the Bulgarian-Polish Bilingual Perspective. – В: Д. Благоева, Т. Александрова (ред.). Доклади от Международната юбилейна конференция на Института за български език „Проф. Любомир Андрейчин“ (София, 2017). Част 1. София: Издателство на БАН „Проф. „Марин Дринов“, 2017, 285-297. (M. Piasecki, S. Koeva. WordNet Relations in the Bulgarian-Polish Bilingual Perspective. – In: D. Blagoeva, T. Aleksandrova (red.). Dokladi ot Mezhdunarodnata yubileyna konferentsiya na Instituta za balgarski eizk “Prof. Lyubomir Andreychin” (Sofia, 2017). Chast 1. Sofia: Izdatelstvo na BAN “Prof. Marin Drinov”, 2017, 285-297.)

Rudnicka et al. 2012: E. Rudnicka, M. Maziarz, M. Piasecki, S. Szpakowicz. A Strategy of mapping plWordNet onto Princeton WordNet. – In: M. Kay, Ch. Boitet (eds). 24th International Conference on Computational Linguistics. Proceedings of COLING 2012: Posters, Mumbai, India, 8-15 December 2012. Mumbai: Indian Institute of Technology, 2012, 1039-1048.

Ruppenhofer et al. 2016: J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, C. F. Baker, J. Scheffczyk. FrameNet II: Extended Theory and Practice, 2016. Available from: https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf [Accessed: 30 March 2021].

Sag et al. 2002: I. Sag, T. Baldwin, F. Bond, A. Copestake, D. Flickinger. Multiword expressions: A pain in the neck for NLP. – Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), 2002, 1-15.

Shi, Mihalcea 2005: L. Shi, R. Mihalcea. Putting pieces together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. – In: A. Gelbukh (ed.). Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science, 3406. Berlin, Heidelberg: Springer, 2005, 100-111.

Tonelli, Pighin 2009: S. Tonelli, D. Pighin. New Features for FrameNet – WordNet Mapping. – Proceedings of the Thirteenth Conference on Computational Natural Language Learning, Boulder, USA, 2009, 455-488.

Tufiş, Cristea, Stamou 2004: D. Tufiş, D. Cristea, S. Stamou. BalkaNet: Aims, methods, results and perspectives: A general overview. – Romanian Journal on Information Science and Technology, D. Tufiş (ed.). Special Issue on BalkaNet, Romanian Academy, 7, 2004, 2-3, 9-43.

Vossen1996: P. Vossen. Right or wrong: Combining lexical resources in the EuroWordNet project. – In: M. Gellerstam, J. Jarborg, S. Malmgren, K. Noren, L. Rogstrom, C. R. Papmehl (eds.). Proceedings of the Euralex Workshop, Göteborg, Sweden, 1996, 715-728.

Vossen et al. 1998: P. Vossen, L. Bloksma, H. Rodriguez, S. Climent, N. Calzolari, A. Roventini, F. Bertagna, A. Alonge, W. Peters. The EuroWordNet Base Concepts and Top-Ontology. Deliverable D017D034D036 EuroWordNet LE2–4003, 1998.

Prof. Svetla Koeva, PhD
Institute for Bulgarian Language "Prof. Lyubomir Andreychin"
Bulgarian Academy of Sciences
52 Shipchenski prohod Blvd., Bl. 17
1113 Sofia, Bulgaria
Email: svetla@dcl.bas.bg