Bulgarian WordNet BulTreeBank WordNet
BTB-WN
Version 4.0

GUIDE

Introduction 1

Representation of parts-of-speech 3

Nouns 3

Adjectives 4

Adverbs 4

Verbs 4

Synset structure and Relations 5

Synset structure 5

Relations 6

Mappings 11**2**

To Open English Wordnet 12

12 13

Mapping with DBpedia

Future work **14**

Sources 13**5**

References 14

Abstract

In this document we present the basic principles and structure of the BulTreeBank WordNet (BTB-WN) for Bulgarian in the context of the decisions that were taken within the time, the construction details as well as the current status of the applied lexical model.

1. Introduction

The creation of BulTreeBank WordNet has a long history. It started initially as Bulgarian domain lexicons aligned to domain and upper ontologies used with the following European projects: LT4EL[1], AsIsKnown[2]. The utility of these lexicons for semantic annotation of domain texts, and some other NLP tasks motivated us to start our own work on Bulgarian WordNet.

BulTreeBank WordNet (BTB-WN) has been created in several steps: (1) by manual translation of English synsets[3] from the Core WordNet subset of Princeton WordNet (5000 more frequently used word senses) into Bulgarian. This step ensures comparable coverage between the two WordNets on the most frequent senses. The translation was done by two people with excellent knowledge of English. First, they formulated a Bulgarian definition reflecting the content of the concept represented by its correspondence to the English synset. Then they formed the Bulgarian synset recording the Bulgarian lemmas that have this meaning. Some of the lemmas might be multiword expressions. After this first phase a lexicographer checked both - the definition and the lemmas. The result from this work was published as part of the Open Multilingual WordNet[4] under CCBY 3.0 licence[5]. This is Version 1.0 of BTB-WN; (2) by identification of senses used in Bulgarian treebank BulTreeBank (BTB). The identified senses have been organised in synsets for the BulTreeBank WordNet. The newly created Bulgarian synsets are being mapped onto the conceptual structure of PWN. In this way, the BTB-WN was extended with real usages of the word meanings in texts. Also, the coverage of the core and base concepts for Princeton WordNet has been evaluated over a Bulgarian syntactic corpus. This is Version 2.0 of BTB-WN. It contained about 9000 synsets; (3) by sense extension, which includes two activities: a) detection of the missing senses of processed lemmas in BulTreeBank and adding them to the BTB-WN, and b) a semi-automatic extraction of information from the Bulgarian Wiktionary mapped to synsets from PWN and then manually checked. After checking a little more than 5000 of them were approved and added in BTB-WN. We would like to thank Antoni Oliver Gonzalez who provided the automatic mapping from Bulgarian Wiktionary to PWN. Behind this extension we added new senses for the words that have been already included in synsets of BTB-WN. The idea is for each word to represent all its senses. This is Version 3.0 of BTB-WN. It contained about 12500 synsets. Both versions were created and used within the QTLeap European project[6]. After the QTLeap project the vocabulary of BTB-WN was extended on the basis of frequency list of lemmas from Bulgarian National Reference Corpus - BulTreeBank. When CLaDA-BG project started at the end of 2018 it contained a little more than 19000 synsets. During CLaDA-BG project 2019, 2020, 2021 and beginning of 2022 BTB-WN was checked by two people for consistency, the definitions were improved, it was mapped to Bulgarian Wikipedia, for many synsets new examples were added.

Thus, the current version of BTB-WN - 4.0 (presented here) - was thoroughly revised using a specified software and extended with more than 10 000 new senses, so currently it contains more than 30 000 synsets.

Several explanatory dictionaries (see section 8) were consulted about the number of word senses in BTB-WN, definitions, etc.

[1] https://cordis.europa.eu/project/id/027391

[2] https://cordis.europa.eu/project/id/028044/it

[3] **Synset** is a structure for the lexical entries in WordNet consisting of a set of synonyms related to the same sense, a definition of the sense, examples of usage of the synonyms.

[4] http://compling.hss.ntu.edu.sg/omw/

[5] https://creativecommons.org/licenses/by/3.0/

[6] https://cordis.europa.eu/project/id/610516

2. Representation of parts-of-speech

BTB-WN currently includes representations of the following four parts of speech: nouns (cardinal numerals, verbal and deverbal nouns are categorised as nouns), adjectives (ordinal numerals are categorised as adjectives; lexicalised participles that function as adjectives are included as well and categorised as adjectives), adverbs and verbs.
Pronouns, prepositions, conjunctions, particles and interjections are considered to be added in the future.

2.1. Nouns

- Grammar
    - Nouns that are defective with respect to the number category (used only in singular (*прах*, “prah”, dust) or only in plural (*анали*, “anali”, annals), always have respectively only singular lemmas and only plural lemmas. The information about their usage is presented in a lemma marker.
- Semantics
    - Professions, roles, titles and ranks for men and women are united in one synset, which has equivalent-to relation with the OEW synset for the noun for men and a near-equivalent-to relation to the OEW synset for women if such is present. Here is an example: *сервитьор*, “servitjor”, waiter and *сервитьорка*, “servitjorka”, waitress are in one synset that has equivalent-to relation with *waiter* and near-equivalent-to with *waitress*.
    - Another approach is taken towards the nouns for male and female relatives and for male and female animals - they belong to different synsets. Here is an example: *баща*, “bašta”, father and *майка* “majka”, mother are in separate synsets and each has equivalent-to relation with respectively *father* and *mother* from OEW. Both *баща* and *майка* have the synset *родител*, “roditel”, parent as a hypernym.
    - Nouns for young animals are presented as synset members of the general meaning of the given animal. Here is an example: *овен*, “oven”, ram, *овца*, “ovca”, sheep and *агне*, “agne”, lamb are in one synset with equivalent-to relation with *sheep* and near-equivalent-to relation with *ram* and *lamb*.
    - Nouns for male and female title and rank holders are members of one synset. Here is an example: *дон*, “don”, Don and *доня*, “donja”, *дона*, “dona”, Dona are in one synset with equivalent-to relation with *Don* and near-equivalent-to relation with *Dona*.
    - Forms of addressing men and women are in one synset. Here is an example: *батко*, “batko” (used for addressing older men) and *кака*, “kaka” (used for addressing older women) are in one synset which has *обръщение*, “obrăštenie” (address) as a hypernym.

      Meanwhile an exception are *господин*, “gospodin”, Mister and *госпожа,* “gospoža”, Mrs. which are in separate synsets and each has equivalent-to relation respectively with *Mr.* and *Mrs.* Both *господин* and *госпожа* have *обръщение* “obrăštenie” (address) as a hypernym.

- Diminutives are members in the synset of the general form of the given word. Here is an example: *стол*, “stol”, chair and *столче*, “stolče”, chair-diminutive are in one synset.

2.2. Adjectives

- Grammar
  In this category are also included participles which function as adjectives under two conditions - a participle is either independently presented in the dictionaries or it is determined as a synonym of an adjective. Here is an example: in the synset with *дебел*, “debel”, fat are added a few participles as synonyms: *охранен*, “ohranen”, *угоен,* “ugoen” respectively from the verbs *охраня,* “ohranya” and *угоя*, “ugoya” (make a person/animal to gain weight with a rich diet), and *хранен*, “hranen” from *храня*, “hranja”, feed. Ordinal numerals (for example, *трети*, “treti”, third) are presented with the adjective category in BTB-WN as it is done in the OEW.
- Semantics
  Both types of Bulgarian adjectives - qualitative (which express intrinsic properties and qualities of an object, for example *красив*, “krasiv”, beautiful) and relative (which reflect qualities and properties of objects in relation to another object, for example *правоъгълен*, “pravoăgălen”, rectangular) - are included in BTB-WN.

2.3. Adverbs

- Grammar
  The two types of Bulgarian adverbs are presented in BTB-WN - regular (derived from nouns, adjectives, numerals, verbs, prepositions, for example *бързо,* bărzo, quickly) and pronominal adverbs (derived from pronouns, for example *тук*, tuk, here).

- Semantics
  Adverbs from all semantic types are included in BTB-WN - qualitative, quantitative, purpose, locative, temporal, etc.

2.4. Verbs

- Grammar
  Impersonal verbs have lemmas in third person singular. Here is an example: *оказва се*, “okazva se”, *окаже се*, “okaže se”, turn out, prove, turn up are in third person singular.
- Semantics
  Prefixed verbs with semantics for beginning, end, duration, etc. of the action are synset members to the general form of the given verb. Here is

an example: *чета*, “četa”, read is in one synset with *зачета*, “začeta”, *зачитам*, “začitam”, start to read, *попрочета*, “poporočeta”, *попрочитам*, “poporočitam”, read a little, partly, *пречета*, “prečeta”, *пречитам*, “prečitam”, read again and so on.

3. Synset structure and Relations

3.1. Synset structure

The synsets in BTB-WN can contain an unlimited number of lemmas, definitions and zero or more example sentences or phrases[7].

The lemmas in BTB-WN can be words, MWEs and abbreviations[8] (in Bulgarian or in English). Synset members can be terms from different language registers and styles (for example: formal, informal, slang, dialect, vulgar, etc.). The specific stylistic features of the synset members are planned to be outlined by a particular lemma order and lemma markers with additional linguistic information.

The following types of additional linguistic information in the form of lemma markers are planned to be introduced:

1. Bulgarian verbs with prefixes that bear semantics of start, end, duration, repeatability, etc. of the action. As mentioned in 4.4., the synsets for these verbs do not have English equivalents, so we decided to unite them with the general forms of the given verb, but define their specifics with lemma markers. For example, the verbs *бягам*, “bjagam”, run and *забягам*, “*zabjagam*”, start to run are members of the same synset. The second, prefixed verb will have a lemma marker that further describes its more specific meaning.

2. As mentioned in 4.1., the diminutive forms of the nouns will be in one synset with the general form. Here is an example: *маса*, “masa”, table and *масичка*, “masička”, table-diminutive are in one synset and *масичка* is labelled with a lemma marker for diminutiveness.

3. Lemma markers will label lemmas from the different language registers and styles - archaic, dialectal, slang, informal, vulgar, offensive, etc. Here is an example: *магданоз*, “magdanoz”, parsley is in the same synset as *мерудия*, “merudija”, *меродия*, “merodija”, parsley-dialectical.

4. There will be a marker for multi-word expression lemmas when they are synset members. Here is an example: *ръчно*, “răčno”, manually is in the same synset as *на ръка*, “na răka”, by hand-MWE.

5. Markers for grammatical specifics will be applied to label:

   a) lemmas of nouns that are mandatory or usually used only in singular or only in plural. Here is an example: *месо*, “meso” and *плът*, “plăt”, flesh-singular. Another example: *хриле*, “hrile”, gill,branchia-plural.

[7] Our goal is all synsets to have examples, but this goal is not completed yet.

[8] Bulgarian synsets can include English abbreviations as lemmas if it is applicable. For example, terms that are used world-wide such as “GUI” (graphical user interface) and “GSM” (Global System for Mobile Communications).

b) lemmas of verbs that are mandatory or usually used only in singular/plural or in specific person. Here is an example: *пия*, “pija”, drink (alcohol) is in the same synset with *изпонапивам се*, “izponapivam se”, *изпонапия се*, “izponapija se”, get drunk (usually used for many or all people, in plural or 3rd person singular).

6. Markers for lemmas that are used only in combination with given word or words. Here is an example: *заседнал*, “zasednal”, stranded, sedentary when used in combination with “life” has the meaning *without active movement, associated with standing still.*

3.2. Relations

1. *Synset-to-synset*

BTB-WN uses all of the relations used in the Open English WordNet and many of the relations from Princeton WordNet and Global WordNet, some of them with modifications.

BTB-WN adopts the concepts for hyponymy and hypernymy relations of PWN, OEW and OMW and considers them as relations between a superordinate and subordinate senses. However, BTB-WN does not completely follow the hypernymy and hyponymy hierarchy of OEW. Additionally, these relations are not used for all PoS and have different importance for them - they are required for the nouns, but not for the verbs; they are not used for adjectives and adverbs.

hypernym

Hypernym relation is used for linking noun and verb synsets. The definition of hypernym relation used in BTB-WN is "a hypernym of something is its superordinate term: if X is a hypernym of Y, then all Y are X" (see further information: https://globalwordnet.github.io/gwadoc/#hypernym).

For example, *ястие*, “jastie”, dish is the hypernym of *супа*, “supa”, soup.

hyponym

Hyponym relation is used for nouns and verbs. "A relation between two concepts where concept B is a type of concept A" is considered a hyponym relation (see further information: https://globalwordnet.github.io/gwadoc/#hyponym). For example *божур*, “božur”, peony is a hyponym of *цвете*, “cvete”, flower.

There are several synset-to-synset relations which have not been purposefully added between Bulgarian synsets, but are present in BTB-WN in the way they are derived from OEW. These relations are planned to be further presented and modified in the next version of BTB-WN.

The *instance_hypernym* and *instance_hyponym* relations are used for nouns and only proper nouns can be instance_hyponyms.

instance_hypernym

This relation indicates "the type of an instance" (see further information - https://globalwordnet.github.io/gwadoc/#instance_hypernym).

For example *столица*, “stolica”, capital is instance_hypernym of *София*, “Sofija”, Sofia.

instance_hyponym

Instance_hyponyms are named entities, for example *Пабло Пикасо*, “Pablo Pikaso”, Pablo Picasso is instance_hyponym of *художник*, “hudožnik”, painter (see further information - https://globalwordnet.github.io/gwadoc/#instance_hyponym).

see also

The *see also* relation, used in OEW and OMW, links two concepts which have a loose semantic relation (for more information - https://globalwordnet.github.io/gwadoc/#also). It is used for nouns, verbs, adjectives and adverbs. For example *безболезнен*, “bezboleznen”, painless has *see also* relation with *безвреден*, “bezvreden”, *безопасен*, “bezopasen” harmless.

attribute

Attribute is a relation between nouns and adjectives where one concept is an attribute of another concept, used in OEW, PWN and OMW (for more information - https://globalwordnet.github.io/gwadoc/#attribute). For example *бавен*, “baven”, slow has an attribute *скорост*, “skorost”, speed, fastness.

causes

The *causes* relation indicates concepts which produce effect or bring a result for another concept and is applied for verbs. It is used in OEW and OMW (for more information - https://globalwordnet.github.io/gwadoc/#causes). For example: *вледеня*, “vledenja”, *вледенявам*, “vledenjavam”, (cause to) freeze is related by causes with *вледя се*, “vledja se”, *вледявам се*, “vledjavam se”, freeze (turn to ice).

is_caused_by

The *is_caused_by* relation is the opposite of causes - it is used for concepts which come as an effect or result from another concept (for more information - https://globalwordnet.github.io/gwadoc/#is_caused_by). For example *страхувам се*, “strahuvam se”, fear, dread has an is_caused_by relation to *плаша*, “plaša”, frighten, scare.

domain_region

The *domain_region* relation indicates a concept which is a geographical or cultural domain for other concepts (for more information - https://globalwordnet.github.io/gwadoc/#domain_region).

For example *гладиатор*, “gladiator”, gladiator has a domain_region *Рим*, “Rim”, Rome.

domain_topic

The *domain_topic* relation is used for concepts which are the scientific category pointer of a given concept (https://globalwordnet.github.io/gwadoc/#domain_topic) and are used in OEW, OMW and PWN (in the latter it is called *domain term category*). For example *бактерия*, “bakterija”, bacteria has a domain_topic *микробиология*, “mikrobiologija”, microbiology.

entails

The *entails* relation indicates verbs which impose another verb as a necessary accompaniment or result and it is used in PWN, OEW and OMW (for more information - https://globalwordnet.github.io/gwadoc/#entails). For example *ям* “jam”, *храня се*, “hranja se”, eat has an entails relation with *дъвча*, “davča”, chew and *гълтам*, “găltam”, swallow.

is_entailed_by

The *is_entailed_by* relation is the opposite of *entails* - it is used for concepts which can not be done unless another concept is done (https://wordnet.princeton.edu/documentation/wngloss7wn). It is applied in OEW and OMW (for more information - https://globalwordnet.github.io/gwadoc/#is_entailed_by). For example, *гълтам*, “galtam”, swallow has an is_entailed_by relation to *храня се*, “hranja se”, eat.

exemplifies

The *exemplifies* relation is used for concepts which are examples other concepts (for more information - https://globalwordnet.github.io/gwadoc/#exemplifies). For example, *без съмнение*, “bez sămnenie”, no doubt has an exemplifies relation with *colloquialism*.

has_domain_region

The *has_domain_region* relation is the opposite of domain_region and it indicates concepts which belong to a given geographical or cultural domain (for more information - https://globalwordnet.github.io/gwadoc/#has_domain_region). It is used in OEW, PWN and OMW. For example *суши*, “suši”, sushi is related with has_domain_region to *Япония*, “Japonia”, Japan.

has_domain_topic

Has_domain_topic is the opposite of domain_topic relation and it indicates a concept which belongs to a given scientific category (for more information - https://globalwordnet.github.io/gwadoc/#has_domain_topic). It is applied in OEW, OMW and in PWN (in the latter it is called *domain category*). For example, *математика*, “matematika”, mathematics has a has_domain_topic relation to *наука*, “nauka”, science.

holo_member
holo_part
holo_substance

Holonym is a concept which has a constituent part, member or substance - meronym. In BTB-WN are used the three types of holonymy relations, which are applied in PWN, OEW and some of the relations used in OMW (for more information - https://globalwordnet.github.io/gwadoc/#holonym):

- holo_member - for holonyms which have members; for example *мит*, “mit”, myth has a

holo_member relation to *митология*, “mitologija”, mythology;
- holo_part - for holonyms which have parts; for example *ноздра*, “nozdra”, nostril has a holo_part relation to *нос*, “nos”, nose;
- holo_substance - for holonyms which have substances (for example *брашно*, “brašno”, flour has a holo_substance relation to *тесто*, “testo”, dough.

mero_member
mero_part
mero_substance
Meronym is a concept which is a constituent part of another concept. In BTB-WN are distinguished the three types of meronyms used in PWN, OEW and some of the relations used in OMW (for more information - https://globalwordnet.github.io/gwadoc/#meronym):
- mero_member - for meronyms that are members of something; for example *галактика*, “galaktika”, galaxy has a mero_member relation to *звезда*, “zvezda”, star;
- mero_part - for meronyms that are part of something; for example *детска площадка*, “detska ploštadka”, playground has a mero_member relation to *люлка*, “ljulka”, swing;
- mero_substance - for meronyms that are the substance of something; for example *тестени изделия*, “testeni izdelija”, pastry has a mero_substace relation to *брашно*, “brašno”, flour. The meronymy relation is suitable only for nouns.

similar
The *similar* relation indicates closely related meanings and in BTB-WN it is applied only for adjectives as it is done in OEW. For example, *необитаем*, “neobitaem”, uninhabited is determined to be *similar* to *изоставен*, “izostaven”, abandoned. The relation is used for more PoS in OMW (for more information - https://globalwordnet.github.io/gwadoc/#similar).

Equivalence relations are used in multilingual wordnets and they are used for identical concepts between two or more languages (Bulgarian and English in the case of BTB-WN). They are suitable for all PoS.

equivalent-to
The *equivalent-to* relation is used between Bulgarian and English synsets for every kind of PoS. It links synsets with exact same meanings. For example, *маса*, “masa”, table has an equivalent_to relation to the synset *table.*

In BTB-WN there are several newly introduced relations between synsets (*near-equivalent-to, sem-derived-from, sem-derives-to, sem-derived-from-v* and *sem-derives-to-v*) and it also adopts some relations from PWN, OEW and OMW which are commonly used in wordnets. The full list of synset-to-synset relations in BTB-WN is:

near-equivalent-to
The *near-equivalent-to* is a recently introduced relation in BTB-WN. It links meanings from all PoS in the two languages which are very similar, but not equal. This relation is very often used to establish a closer connection between a head adjective and its satellites - frequently a

Bulgarian adjective is an exact equivalent of the OEW head adjective but its satellite adjectives are semantically too similar so they become near-equivalents to the Bulgarian adjective. For example: *незнаен, "neznaen", unknown* ('Which is not known') has *equivalent-to* relation with *unknown* ('not known') and *near-equivalent* with *unfamiliar* ('not known or well known').

In BTB-WN there are four relations (*sem-derived to* and *from*) which indicate semantic relatedness between adjectives and nouns and adjectives and verbs, but they also have derivational relation of varying kinds. These relations currently are not strictly following the derivational paradigm, they do not reflect the word formation direction - the semantic connection between these PoS is in the first place.

Precise derivational relations are part of the plan for future versions of the wordnet. For example, the adjective *радостен*, "radosten*"*, happy has sem-derived-from relations both with the noun *радост,* "radost*"*, joy, happiness and the verb *радвам се*, "radvam se*"*, rejoice, joy.

sem-derived-from

The *sem-derived-from* relation determines adjectives that are semantically and derivationally related with a noun. The relation is particularly useful for adjectives which do not have OEW equivalents. Since the *similar* and *antonymy* relation has not been systematically used in BTB-WN, this type of adjectives would not have any link with the hierarchy. For example, *диамантен,* "diamanten" (adjective from *diamond*) has a sem-derived-from relation to *диамант*, "diamant", diamond.

Sem-derived-from would be beneficial also to establish relations between adverbs and nouns. *Sem-derived-from* could be also used between adverbs and nouns.

sem-derives-to

The *sem-derives-to* relation is used for nouns which are semantically and derivationally related with adjectives. For example, *дантела*, "dantela", lace has a sem-derives-to relation to *дантелен*, "dantelen" (adjective from *lace*).

sem-derived-from-v

Sem-derived-from-v relation is used for adjectives that are semantically and derivationally related with a verb. For example*, административен*, "administrativen", administrative has a sem-derived-from-v relation to *администрирам*, "administriram", administrate.

sem-derives-to-v

The *sem-derives-to-v* relation is used for verbs and it indicates the adjectives with which they are related. For example, *администрирам*, "administriram", administrate has a sem-derives-to-v relation to *административен*, "administrativen", administrative.

sem-derived-from-adj

The relation *sem-derived-from-adj* is used for adverbs, which are related with adjectives. For example, тъмно, "tămno", darkly has a *sem-derived-from-adj* relation with тъмен, "tămen", dark.

sem-derives-to-adj
The relation *sem-derives-to-adj* shows with which adverbs an adjective is related. For example, *артистичен*, "artističen", artistic has a *sem-derives-to-adj* relation with *артистично,* "aritistično", artistically.

predecessor
The relation *predecessor* is used only for named entities and connects a concept with its successor. For example: *Университет за национално и международно стопанство*, "Universitet za nacionalno i meždunarodno stopanstvo", University of national and world economy" has a relation *predecessor* with *Висш икономически институт „Карл Маркс"*, "Visš ikonomičeski institut "Karl Marks", Karl Marx Higher Institute of Economics.

successor
The relation *successor* is used only for named entities and connects a concept with its predecessor. For example: *Свободен университет за политически и стопански науки,* "Svoboden universitet za političeski i stopanski nauki", Free university for political and economic sciences has a relation *successor* with *Държавно висше училище за финансови и административни науки*, "Dăržavno visše učilište za finansovi i administrativni nauki, State higher school of financial and administrative sciences.

2. *Lemma-to-lemma*

In BTB-WN there are not any relations between lemmas, the main focus in the current version is on the synset-to-synset relations. The introduction of lemma-to-lemma relations is considered as a part of the future work on the wordnet. There will be at least two relations of this type - antonymy and derivational relation.

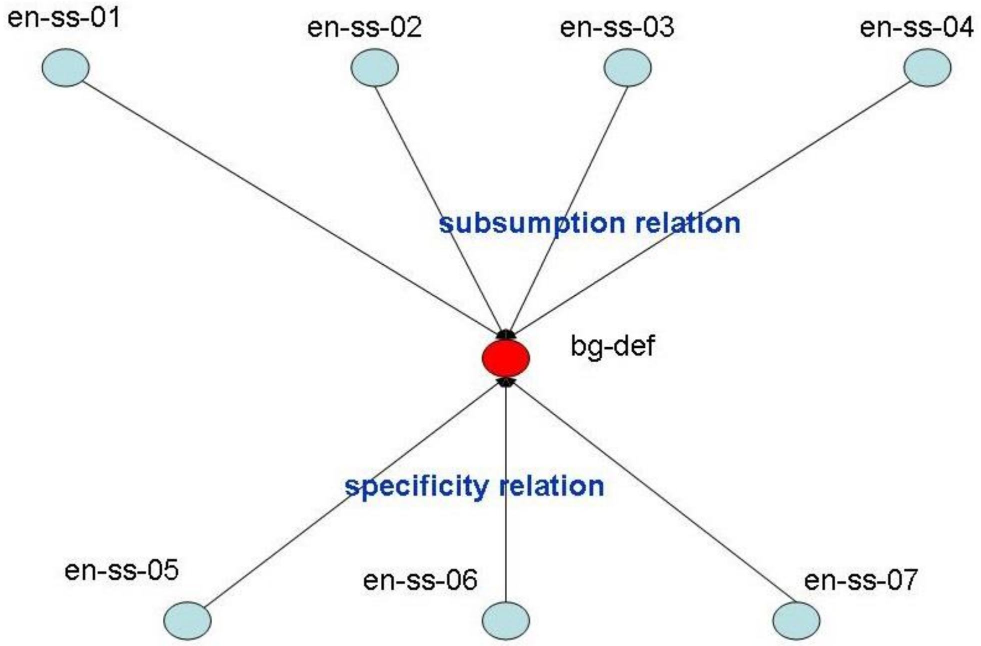
6. Mappings

6.1. To Open English Wordnet (OEW)

In this section we present the mapping that we established to OEW (initially to Princeton Wordnet 3.0).
The process of mapping word senses to WordNet is concept-based, i.e., mapping concepts takes precedence over mapping words. For instance, a Bulgarian word may be mapped to an English one, in accordance with the dictionary, but the English word may not refer to the concept in question. In cases such as these, the mapping is done via strategies other than mapping word-to-word (see section 3 above).

BTB-WN is mapped first with the Princeton WordNet 3.0 and later also with the OEW. First, a manual translation of the Core WordNet subset of Princeton WordNet in Bulgarian was done with the purpose of having a comparable coverage between the two resources. The second step includes the identification of senses found in the Bulgarian treebank BulTreeBank, and the organisation of these senses in Bulgarian synsets, which are then mapped to the conceptual structure of PWN. This process ensures the presence of real usages of senses in texts. The mapping process starts with translation of a Bulgarian term in English, then search for the corresponding English synset and establishment of relation between the two synsets (the relations used in BTB-WN are introduced chapter 5 - Synset structure and Relations), addition of Bulgarian definition and examples.

Since 2020 BTB-WN is mapped to the OEW and the main benefits of the mapping are that this wordnet is being updated, edited and expanded (unlike PWN). OEW is connected with Wikidata and Wikipedia, so it could be used for knowledge transfer to BTB-WN.



6.2. Mapping with Wikipedia

Two types of extension of BTB-WN were intended - extension of the existing lemmas with new senses and extension with instances. For the first task all the lemmas in BTB-WN were compared with the titles of Bulgarian Wikipedia articles and the senses from Wikipedia, which were missing in BTB-WN, were added in the wordnet with definitions and links to Wikipedia. The titles of the corresponding English Wikipedia articles were also extracted and used for the selection of right sense in English and thus, an appropriate synset in OEW for mapping. The disambiguation pages in Wikipedia proved useful, because they include information about related words. For this mapping three types of correspondences between BTB-WN and Bulgarian Wikipedia were distinguished - equality of the senses in the two resources, concept available in Wikipedia, but missing in BTB-WN and NE represented in Wikipedia, but not in the wordnet. In the last two cases a new synset was created and mapping was established also with PWN.

6.3. Mapping with DBpedia

The mapping with DBpedia[9] (an open knowledge graph with information from Wikimedia) was used for the second type of BTB-WN extension - with instances. Named entities in BulTreeBank are annotated with URIs from DBpedia and because Bulgarian DBpedia is relatively small, Bulgarian Wikipedia was also used. The NEs, which were missing in Wikipedia and DBpedia were annotated with the classes of DBpedia ontology. This way the instances of the classes were automatically classified as instances of the corresponding synsets. So far, the mapping is done with the three most frequent types of Named entities in the BulTreeBank - people, locations and organisations.

This mapping is also partially supported by the mapping to Bulgarian Wikipedia. It is thus applicable for other knowledge bases originated from Wikipedia.

7. Future work

The extensions of BTB-WN will be done in several directions. First, we will constantly extend its vocabulary. The new entries will be not only from common language, but also terms forming various domains. We are planning to map different ontologies to BTB-WN in order to support language access to knowledge bases constructed with respect to these ontologies.

One task on which we are already working is a mapping from BTB-WN to a Valency lexicon of Bulgarian and a mapping with Predicate Matrix. The former one is part of our in-house plan for integrating all Bulgarian resources, while the latter task would test the level of automatic knowledge transferability between English and Bulgarian with a post-editing step. It is relevant because English remains the language with more and better integrated language resources. As an initial probe, the texts from the parallel English-Bulgarian parts of the SETimes corpus[10] were annotated - first, the English part of them was annotated with Predicate Matrix sense and then these senses have been transferred to the Bulgarian part. In the transferred senses several cases were observed: (1) the Bulgarian sense differs from the one in Predicate Matrix, but still is valid; (2) several senses are matched including the correct one; (3) one sense is matched from BTB-WN and it is the correct one; (4) wrong sense is mapped since the corresponding sense of the lemma is missing in BTB-WN. There were also non-transferred senses because of errors in the part-of-speech tagger (English or Bulgarian) or a missing lemma in BTB-WN. Thus, new experiments are envisaged here which would use a BTB-WN with better coverage.

[9] https://www.dbpedia.org/

[10] http://nlp.ffzg.hr/resources/corpora/setimes/

8. Sources

Here we list the main sources with which we consult during the creation of BTB-WN in order to do the determination of the possible senses, the formulation of definitions, examples, mapping to English Princeton WordNet and to English Open WordNet.

1. Multivolume dictionary of Bulgarian language (https://ibl.bas.bg/rbe/)
2. L. Andrejčin, et al. Bulgarian explanatory dictionary, IV edition, supplemented and revised by D. Popov. Nauka i izkustvo, 1994
3. E. Perniška, D. Blagoeva and S. Kolkovska. Dictionary of the new words in Bulgarian language, Nauka i izkustvo, 2010, 2021
4. D. Blagoeva and S. Kolkovska. Dictionary of the new words in Bulgarian language, Nauka i izkustvo, 2021
5. I. Kasabov and K. Stojanov. Universal encyclopaedic dictionary, Svidas, 1999, 2003
6. A. Nanova. Bulgarian synonymy and antonymy dictionary with idioms, Prosveta, 2019
7. Online dictionary (http://www.onlinerechnik.com/)
8. English Princeton WordNet - https://wordnet.princeton.edu/
9. English Open WordNet - https://en-word.net/
10. Wikipedia - https://bg.wikipedia.org/wiki/Начална_страница
11. Wiktionary - https://bg.wiktionary.org/wiki/Уикиречник:Начална_страница

In the process of the creation of BTB-WN we also consult the concordances over several Bulgarian corpora:

1. Bulgarian HPSG-based TreeBank
2. Bulgarian National Reference Corpus - BulTreeBank
3. CLaDA-BG Multi Billion Corpus

9. References

1. Laskova et al. 2019: Laska Laskova, Petya Osenova, Kiril Simov, Ivajlo Radev, Zara Kancheva 2019: *Modeling MWEs in BTB-WN*. In Agata Savary, Carla Parra Escartín, Francis Bond, Jelena Mitrović, Verginica Barbu Mititelu (eds.) Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). Association for Computational Linguistics, pp. 70-78. ISBN 978-1-950737-26-0
2. Osenova et al. 2017: Petya Osenova and Kiril Simov 2017: *Challenges Behind the Data-driven Bulgarian WordNet (BulTreeBank Bulgarian Wordnet)*. In: John P. McCrae, Francis Bond, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Jorge Gracia, Ilan Kernerman, Elena Montiel Ponsoda, Noam Ordan and Maciej Piasecki (eds), Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets, co-located with 1st Conference on Language, Data and Knowledge (LDK 2017, Galway, Ireland, June 18, 2017,

Vol-1899 urn:nbn:de:0074-1899-7, pp.152-163
(http://ceur-ws.org/Vol-1899/)(http://ceur-ws.org/Vol-1899/CfWNs_2017_proc4-paper_3.pdf)

3. Simov et. al 2019: Kiril Simov, Petya Osenova, Laska Laskova, Ivajlo Radev, Zara Kancheva 2019: *Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia.* In: Fellbaum, Christiane and Vossen, Piek and Rudnicka, Ewa and Maziarz, Marek and Piasecki, Maciej (eds) Proceedings of the Tenth Global Wordnet Conference, pp. 290-297. ISBN 978-83-7493-108-3
4. Osenova, P. and Simov, K. (2018). The data-driven Bulgarian WordNet: BTB-WN. Cognitive Studies Études cognitives, 2018 (18). https://doi.org/10.11649/cs.1713
5. John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – An Open-Source WordNet for English. In Proceedings of the 10th Global Wordnet Conference, pages 245–252, Wroclaw, Poland. Global Wordnet Association.
6. Popov, A., Kancheva, S., Manova, S., Radev, I., Simov, K., & Osenova, P. (2014). The sense annotation of BulTreeBank. In V. Henrich, E. Hinrichs, D. de Kok, P. Osenova, & A. Przepiorkowski (Eds.), Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13) (pp. 127–136)