

Българската мрежа от думи
(BulTreeBank WordNet - BTV-WN)
Версия 4.0

ИСТОРИЯ НА РЕСУРСА И РЪКОВОДСТВО

Увод	1
Представяне на части на речта	3
Съществителни имена	3
Прилагателни имена	4
Наречия	5
Глаголи	5
Структура на синонимните гнезда и релации	5
Структура на синонимните гнезда	5
Релации	6
Съответствия	12
Съответствие със Свободната английска мрежа от думи (Open English Wordnet)	12
Съответствие с Уикипедия	13
Съответствие с DBpedia	13
Бъдеща работа	15
Източници	146
Библиография	156

Резюме

В този документ са представени основните принципи и структурата на BulTreeBank WordNet (BTV-WN) за български език в контекста на решенията, които са взети през времето, детайлите за конструирането му, както и текущото състояние на приложения лексикален модел.

1. Увод

Създаването на *Българската Мрежа от думи (BTV-WN)* има около 18-годишна история. През 2005 г. във връзка с участието на [групата Бултрибанк](#) в редица европейски проекти се появява нуждата от свързване на онтологични (езиково независими) понятия

в дадена тематична област с лингвистичните термини (езиково специфични термини, ключови думи, лексика) в различни езици. Затова работата на групата започва с набор от български терминологични речници, подравнени със специализирани и общи онтологии, както и със съществуващия вече Onto WordNet. Те се използват в следните европейски проекти: LT4EL¹ (в областта на компютърните умения), AsIsKnown² (в областта на вътрешния дизайн), QTLeap³ (в областта на използването на софтуер) и др. Използването на тези речници в семантичната анотация на текстове от съответните тематични области, в машинния превод, извличането на знание, обучението по граматика и др. ни мотивира да започнем работата по създаване на свободно достъпна *Българска мрежа от думи* (BulTreeBank WordNet (BTB-WN)).

BTB-WN е създадена чрез няколко стъпки: (1) чрез ръчен превод на английски синонимни гнезда⁴ от подгрупата с емпирично избрани основни понятия Core WordNet от английската мрежа от думи Princeton WordNet (PWN) (5000 по-често използвани значения) на български език. Тази стъпка осигурява подходящо покритие между двете мрежи от думи за най-често срещаните значения. Преводът е направен от двама души с отлични познания по английски. Те са процедирали по следния начин: първо, формулирали са българската дефиниция, която отразява съдържанието на понятието, представено от съответствието му с английското синонимно гнездо. След това са оформили българското синонимно гнездо, като са записали българските леми, които имат това значение. Някои от лемите са многокомпонентни думи (Multiword expressions (MWE)). След първата фаза опитен лексикограф е проверил дефиницията и свързаните с нея леми.

Резултатът от тази работа е публикуван като част от многоезичната мрежа от думи Свободна многоезикова мрежа от думи (Open Multilingual WordNet⁵ (OMW)) с лиценз CC BY 3.0⁶. *Това е версия 1.0 на BTB-WN*; (2) чрез идентифициране на значения, използвани в Българския корпус със синтактични анализи - BulTreeBank (BTB). Разпознатите значения са организирани в синонимни гнезда за BTB-WN. Новосъздадените български синонимни гнезда са съотнесени с понятийната структура на PWN. По този начин BTB-WN е разширена с реални употреби на значенията в текстове. Също така покритието на основните понятия от Princeton WordNet в българския език е изчислено на базата на българския синтактичен корпус BulTreeBank. *Това е версия 2.0 на BTB-WN*.

Тя съдържа около 9000 синонимни гнезда; (3) чрез разширяване на значенията, което включва две дейности: а) разпознаване на липсващи значения на обработени леми в BulTreeBank и добавянето им в BTB-WN и б) полуавтоматично извличане на информация от българския Уикиречник и съотнасяне на тази информация със синонимни гнезда от PWN. По-късно е направена ръчна проверка. След проверката

¹ <https://cordis.europa.eu/project/id/027391>

² <https://cordis.europa.eu/project/id/028044/it>

³ <http://qt leap.eu/>

⁴ Синонимното гнездо е структура за лексикалните елементи в уърднет, която се състои от сбор от синоними, свързани със същото значение, дефиниция на значението, примери с употреби на синонимите.

⁵ <http://compling.hss.ntu.edu.sg/omw/>

⁶ <https://creativecommons.org/licenses/by/3.0/>

повече от 5000 от синонимните гнезда са одобрени и добавени в ВТВ-WN. Искаме да благодарим на Антони Оливър Гонзалез, който осигури автоматичното съотнасяне на българския Уикиречник с PWN. След това разширяване бяха добавени нови значения на думите, като те вече са включени в синонимните гнезда на ВТВ-WN. Целта е била да са представени всички значения на дадена дума. *Това е версия 3.0 на ВТВ-WN*. Тя съдържа около 12500 синонимни гнезда. И двете версии (2.0 и 3.0) са създадени и използвани в рамките на проекта QTLear European project⁷. След края на проекта речникът на ВТВ-WN е разширен на базата на честотен списък с лема от Българския национален референтен корпус - BulTreeBank. Когато проектът CLaDA-BG започна в края на 2018 г., ВТВ-WN съдържаше малко повече от 19000 синонимни гнезда. По време на проекта CLaDA-BG през 2019 г., 2020 г., 2021 г. и в началото на 2022 г. ВТВ-WN е проверявана за качество от поне двама души, като са били подобрявани дефинициите; правени са съответствия с българската Уикипедия, добавяни са нови примери за много синонимни гнезда.

Текущата версия на ВТВ-WN - 4.0 (която е представена тук) - е подробно проверена с помощта на специализиран софтуер и разширена с повече от 10 000 нови значения, така че в момента мрежата от думи съдържа повече от 30 000 синонимни гнезда. Няколко тълковни речника (виж раздел 8) са използвани за справка на броя значения в ВТВ-WN, за проверка на дефиниции и други.

2. Представяне на включените части на речта

В момента ВТВ-WN включва представяния на следните четири части на речта: съществителни (където влизат бройните числителни имена и отглаголните съществителни имена), прилагателните имена (където влизат редните числителни имена и лексикализираните причастия, които функционират като прилагателни), наречия и глаголи.

Местоимения, предлози, съюзи, частици и междуметия ще бъдат добавени в бъдеще.

2.1. Съществителни имена

- Граматика
 - Съществителни имена, които се използват само в единствено (*прах*) или само в множествено число (*анали*), винаги имат лема съответно само за единствено или множествено число. Информацията за тяхната употреба ще бъде представена в допълнителен маркер към лемата.
- Семантика
 - Професии, роли, титли и рангове за мъже и жени са обединени в едно синонимно гнездо, което има релацията *равенство-със* (equivalent-to) със

⁷ <https://cordis.europa.eu/project/id/610516>

синонимното гнездо от свободната английска мрежа от думи Open English WordNet (OEW) за съществителното за мъж и релацията *почти-равенство-със* (near-equivalent-to) с английското синонимно гнездо за жена, ако такова е налично. Например: *сервитьор* и *сервитьорка* са в едно синонимно гнездо и имат equivalent-to релация с английското синонимно гнездо за *waiter* и near-equivalent-to релация с *waitress*.

- Различен подход е приет за съществителните за мъже и жени роднини и мъжки и женски животни - те принадлежат на отделни синонимни гнезда. Например: *баща* и *майка* са в отделни синонимни гнезда, баща има equivalent-to релация към *father*, а майка - към *mother* от OEW. *Баща* и *майка* имат еднакъв хипероним - синонимното гнездо *родител*.
- Съществителните имена за малки животни са представени като членове на синонимното гнездо за основното значение на дадено животно. Например: *овен*, *овца* и *агне* са в едно синонимно гнездо с equivalent-to релация към *sheep* и near-equivalent-to релация към *ram* и *lamb*.
- Съществителните за мъже и жени носители на титли и рангове са членове на едно синонимно гнездо. Например: *дон* и *доня* са в едно синонимно гнездо с equivalent-to релация към *Don* и near-equivalent-to *Dona*.
- Обръщенията за мъже и жени са в едно синонимно гнездо. Например: *батко* (за обръщение към по-възрастен мъж) и *кака* (за обръщение към по-възрастна жена) са в едно синонимно гнездо, което има хипероним *обръщение*. Изключение от това правило са *господин* и *госпожа*, които са в отделни синонимни гнезда и всяко има equivalent-to релация съответно към *Mister* и *Mrs*.
- Умалителните форми са членове на синонимното гнездо на основната форма на думата. Например: *стол* и *столче* са в едно синонимно гнездо.

2.2. Прилагателни имена

- Граматика

В тази категория също се включват и причастия, които функционират като прилагателни имена, при две условия - причастията има речникова статия в речниците или е определено като синоним на прилагателно име. Например: прилагателното *дебел* има няколко синонима причастия - *охранен* (от глагола *охранвам*), *угоен* (от *угоявам*) и *хранен* (от *храня*). Редните числителни имена (например *трети*) са причислени към категорията на прилагателните в BTB-WN, както и в OEW.

- Семантика

Двата вида български прилагателни имена - качествени (които изразяват вътрешни свойства и качества на обект, например *красив*) и относителни (които отразяват качества и свойства на обекти по отношение на друг обект, например *правоъгълен*) са включени в BTB-WN.

2.3. Наречия

- Граматика

Двата вида български наречия са представени в ВТВ-WN - обикновени (образувани от съществителни, прилагателни, числителни, глаголи, предлози, например *бързо*) и местоименни наречия (образувани от местоимения, например *тук*).

- Семантика

Всички семантични видове наречия са включени в ВТВ-WN - качествени, количествени, за цел, за място, за време и други.

2.4. Глаголи

- Граматика

Безличните глаголи имат леми в трето лице, единствено число. Например: *оказва се, окаже се*.

- Семантика

Глаголите с представки със значения за започване, край, продължителност и т. н. на действието са членове на синонимното гнездо на основната форма на дадения глагол. Например: *чета* е в едно синонимно гнездо със *зачитам* (със значение “започвам да чета”), *попрочитам* (“чета малко”), *пречитам* (“чета отново”) и други.

3. Структура на синонимните гнезда и релации

3.1. Структура на синонимните гнезда

Синонимните гнезда в ВТВ-WN съдържат дефиниция, неограничен брой леми и могат да имат нула или повече примерни изречения или фрази⁸.

Лемите в ВТВ-WN могат да бъдат думи, многокомпонентни думи и аббревиатури⁹ (на български или на английски език).

Членове на синонимните гнезда могат да бъдат термини от различните езикови регистри и стилове (например: официален, неофициален, сленг, диалектен, вулгарен и т. н.). Специфичните стилистични характеристики на членовете на синонимните гнезда ще бъдат обозначени с определен ред на лемите и с маркери за лема, съдържащи допълнителна лингвистична информация.

⁸ Нашата цел е всички синонимни гнезда да имат примери, но това все още не е осъществено.

⁹ Българските синсети могат да имат леми английски аббревиатури, когато е приложимо. Например термини, които се използват в целия свят, като “GUI” (graphical user interface) и “GSM” (Global System for Mobile Communications).

Планирано е да се въведат следните видове допълнителна лингвистична информация под формата на маркери за лема:

1. Българските глаголи с представки, които имат значение за начало, край, продължителност, повторемост и други на действието. Както е посочено в 4.4., синонимните гнезда на тези глаголи нямат английски еквиваленти, затова решихме да ги обединим с основната форма на дадения глагол, но да отбележим техните особености с маркери за лема. Например: глаголите *бягам* и *забягам* (в значение “започвам да бягам”) са в едно синонимно гнездо. Глаголът с представка *забягам* ще има маркер за лема, който описва специфичното му значение.
2. Както е посочено в 4.1., умалителните форми на съществителните имена са в едно синонимно гнездо с основната форма. Например: *маса* и *масичка* са в едно синонимно гнездо и *масичка* ще е отбелязана с маркер за лема за умалителност.
3. Маркери за лема ще отбелязват лемите от различни езикови регистри и стилове - архаични, диалектни, сленг, вулгарни, обидни и други. Например: *магданоз* е в едно синонимно гнездо с *мерудия* (диалектно название на магданоза).
4. Ще има маркер за лема за многокомпонентните думи, например: *ръчно* е в едно синонимно гнездо с многокомпонентната дума *на ръка*.
5. Маркери за граматични особености ще се прилагат за да отбелязват:
 - а) лемите на съществителни, които задължително или обикновено се използват само в единствено или само в множествено число. Например: *месо*, *плът* (винаги в единствено число) или *хриле* (винаги в множествено число).
 - б) лемите на глаголи, които задължително и обикновено се използват само в единствено/множествено число или в определено лице. Например: *пия* е в едно синонимно гнездо с *изпонапивам се*, което обикновено се използва в трето лице единствено число (“*персоналът се изпонапи и изчезна*”, “*цялото племе взе да се бие, като се изпонапи*”) или в множествено число (“*едни младежи се изпонапиха и взеха да разпалват голям огън*”, “*мъжете са се изпонапили от мъка*”).
6. Маркери за лемите, които се използват само в комбинация с друга дума или думи. Например *заседнал* в комбинация с *живот* означава “живот без активно движение, свързан със стоене на едно място”.

3.2. Релации

1. Между синонимни гнезда

ВТВ-WN използва всички релации, които се използват в Open English WordNet и много от релациите в Princeton WordNet и Global WordNet, някои от тях с модификации. ВТВ-WN възприема понятията за хипонимни и хиперонимни релации на PWN, OEW and OMW и ги смята за отношения между подпонятия и надпонятия, но не следва йерархията на OEW напълно. Тези релации не се използват за всички части на речта и имат различна важност за тях - изискват се за съществителните имена, но не и за глаголите; не се

използват за прилагателни и наречия.

hypernym

Релацията *хипероним* (*hypernym*) се използва за свързване на синонимни гнезда на съществителни и глаголи. Дефиницията за хиперонимната релация, използвана в ВТВ-WN, е “хипероним на нещо е надпонятието на съответното понятие: ако X е хипероним на Y, значи всички Y са X” (виж повече информация: <https://globalwordnet.github.io/gwadoc/#hypernym>). Например: *ястие* е хипероним на *супа*.

hyponym

Релацията *хипоним* (*hyponym*) се използва за съществителни и глаголи. “Релация между понятия, при която понятие B е вид понятие A” се смята за хипонимна релация (виж повече информация: <https://globalwordnet.github.io/gwadoc/#hyponym>).

Например: *божур* е хипоним на *цвете*.

Има няколко релации между синонимни гнезда, които не са целенасочено и последователно добавени между българските синонимни гнезда, но са налични в ВТВ-WN във вида, в който са наследени от OEWN. Планирано е тези релации да се представят системно и да се модифицират в следващата версия на ВТВ-WN.

Релацията за отделен представител на хипероним (*instance_hyponym*) и за отделен представител на хипоним (*instance_hyponym*) се използват за съществителни имена, като само собствените съществителни могат да бъдат *instance_hyponym*.

instance_hyponym

Релацията отбелязва типа на отделния представител ("the type of an instance") (виж повече информация: https://globalwordnet.github.io/gwadoc/#instance_hyponym). Например: *столица* е *instance_hyponym* на *София*.

instance_hyponym

Синонимните гнезда с релация *instance_hyponym* са именувани същности (виж повече информация: https://globalwordnet.github.io/gwadoc/#instance_hyponym). Например: *Пабло Пикасо* е *instance_hyponym* на *художник*.

see also

Релацията *виж още* (*see also*), която се използва в EWN и OMW, свързва понятия със свободна семантична връзка (виж повече информация: <https://globalwordnet.github.io/gwadoc/#also>). Тя се използва за съществителни, глаголи, прилагателни и наречия. Например: *безболезнен* има *see also* релация с *безвреден* и *безопасен*.

attribute

Атрибут (*attribute*) е релация между съществителни и прилагателни, при която едното

понятие е атрибут на другото понятие. Използва се в EWN, PWN and OMW (виж повече информация: <https://globalwordnet.github.io/gwadoc/#attribute>).

Например: *бавен* има *attribute* *скорост*.

causes

Релацията *причинител* (*causes*) отбелязва понятия, които произвеждат някакъв ефект или водят до резултат за друго понятие, и се използва в EWN and OMW за глаголи (виж повече информация: <https://globalwordnet.github.io/gwadoc/#causes>).

Например: *вледенявам*, *вледеня* (със значение “правя нещо да се превърне в лед или да се покрие с лед”) има релация *causes* с *вледявам се*, *вледя се* (със значение “превръщам се в лед”).

is_caused_by

Релацията *причинено от* (*is_caused_by*) е противоположна на релацията *причинител* (*causes*) - тя се използва за понятия, които се случват като ефект или в резултат от друго понятие (виж повече информация: https://globalwordnet.github.io/gwadoc/#is_caused_by).

Например: *страхувам се* има *is_caused_by* релация с *плаша*.

domain_region

Релацията *област на употреба* (*domain_region*) отбелязва понятие, което е географска или културна област за други понятия (виж повече информация: https://globalwordnet.github.io/gwadoc/#domain_region). Например: *гладиатор* има *domain_region* релация с *Рим*.

domain_topic

Релацията *тема в област* (*domain_topic*) се прилага за понятия, които посочват научната категория на дадено понятие. Релацията се използва в EWN, OMW and PWN (в последния се нарича *domain term category*). Например: *бактерия* има *domain_topic* релация с *микробиология*.

entails

Релацията *от мен следва нещо* показва глаголи, които налагат друг глагол да е в комбинация с тях или да се явява като резултат, и се използва в PWN, EWN and OMW (виж повече информация: <https://globalwordnet.github.io/gwadoc/#entails>). Например: *ям* има *entails* релация с *дъвча* и с *гълтам*. Т.е. Ако ям, то аз дъвча и гълтам.

is_entailed_by

Релацията *аз следвам от нещо* е противоположна на релацията *от мен следва нещо* - използва се за понятия, които могат да бъдат осъществени, само ако друго понятие е осъществено също (<https://wordnet.princeton.edu/documentation/wngloss7wn>). Използва се в EWN и OMW (виж повече информация:

https://globalwordnet.github.io/gwadoc/#is_entailed_by).

Например: *гълтам* има *is_entailed_by* релация с *храня се*.

exemplifies

Релацията *давам_като_пример* (*exemplifies*) се използва за понятия, които са примери за други понятия (виж повече информация:

<https://globalwordnet.github.io/gwadoc/#exemplifies>). Например: *no doubt* (без съмнение) има релация *exemplifies* с *colloquialism* (разговорна дума).

has_domain_region

Релацията *имам_област_на_употреба* (*has_domain_region*) е противоположната на *област на употреба* (*domain_region*) и отбелязва понятия, които принадлежат към определена географска или културна област (виж повече информация: https://globalwordnet.github.io/gwadoc/#has_domain_region). Използва се в EWN, PWN and OMW. Например: *суши* има *has_domain_region* релация с *Япония*.

has_domain_topic

Релацията *има_научна_област/има_тематична_област* (*has_domain_topic*) е противоположна на *тема_в_област* (*domain_topic*) (виж повече информация тук: https://globalwordnet.github.io/gwadoc/#has_domain_topic). Използва се в EWN, OMW и PWN (в последния се нарича *domain category*). Например: *математика* има *has_domain_topic* релация с *наука*.

holo_member

holo_part

holo_substance

Холоним е понятие, което има мероним - съставна част, член или субстанция. В ВTB-WN се използват три вида холонимни релации, които се прилагат и в PWN, EWN, а някои от тях и в OMW (виж повече информация:

<https://globalwordnet.github.io/gwadoc/#holonym>):

- *холо_член* (*holo_member*) - за холоними, които имат членове. Например: *мит* има *holo_member* релация с *митология*;

- *холо_част* (*holo_part*) - за холоними, които имат части. Например: *ноздра* има *holo_part* релация с *нос*;

- *холо_субстанция* (*holo_substance*) - за холоними, които имат субстанция. Например: *брашно* има *holo_substance* релация с *тесто*.

mero_member

mero_part

mero_substance

Мероним е понятие, което е съставна част от друго понятие. В ВTB-WN се разграничават три вида мероними, които се използват в PWN и EWN, а някои от тях и в OMW (виж повече информация: <https://globalwordnet.github.io/gwadoc/#meronym>):

- *меро_член* (*mero_member*) - за мероними, които са членове на нещо. Например:

- галактика* има *mero_member* релация със *звезда*;

- *меро_част* (*mero_part*) - за мероними, които са част от нещо. Например: *детска*

площадка има *mero_part* релация с *люлка*;

- *mero_субстанция* (*mero_substance*) - за мероними, които са субстанция на нещо.

Например: *хляб* има *mero_substance* релация с *брашно*.

Меронимната релация е подходяща само за съществителни имена.

similar

Релацията *подобност* (*similar*) посочва тясно свързани понятия и в ВTB-WN, както и в ОEW, се прилага само за прилагателни имена. Например: *необитаем* има релация *similar* с *изоставен*. Релацията се използва за повече части на речта в ОMW (виж повече информация: <https://globalwordnet.github.io/gwadoc/#similar>).

Еквивалентни релации се използват в многоезичните мрежи от думи за свързване на равностойни понятия от два или повече езици (български и английски в ВTB-WN). Те са подходящи за всички части на речта.

equivalent-to

Релацията *равенство-с* (*equivalent-to*) се използва между български и английски синонимни гнезда за всички части на речта. Тя свързва синонимни гнезда с напълно равностойни значения. Например: *маса* има *equivalent_to* релация с *table*.

В ВTB-WN има няколко нововъведени релации между синонимни гнезда:

near-equivalent-to

Релацията *почти_равенство* (*near-equivalent-to*) е наскоро въведена в ВTB-WN. Тя свързва значения от всички части на речта между двата езика, които са много близки, но не и напълно равностойни. Релацията често се използва, за да установи близка връзка между главно прилагателно (*head adjective*) и неговите сателити (*satellites*) - често българско прилагателно е точен еквивалент на английското главно прилагателно, но неговите сателитни прилагателни са прекалено близки по семантика и получават релация *почти_равенство* с българското прилагателно. Например: *незнаен* ('Който, когото не знаем или не сме чували за него, за който няма сведения.') има *equivalent-to* релация с *unknown* ('not known') и *near-equivalent* с *unfamiliar* ('not known or well known').

В ВTB-WN има четири релации (*sem-derived-from* и *sem-derives-to*), които посочват семантична свързаност между прилагателни и съществителни имена, прилагателни и глаголи и наречия и прилагателни, но и различни видове словообразователни връзки. Тези релации към момента не следват стриктно словообразователната парадигма, не отразяват посоката на словообразуване - семантичната връзка между тези части на речта е на първо място.

Прецизните словообразователни релации са част от плана за следващите версии на мрежата от думи. Например: прилагателното *радостен* има *sem-derived-from* релации със съществително *радост* и с глагола *радвам се*.

sem-derived-from

Релацията *образуван_от* (*sem-derived-from*) отбелязва прилагателни, които са семантично и словообразователно свързани със съществително име. Релацията е особено полезна за прилагателни, които нямат английски еквиваленти. Тъй като релациите *подобност* (*similar*) и *антонимия* (*antonymy*) не са систематично използвани в ВТВ-WN, този тип прилагателни не биха имали никаква връзка с йерархията. Например: *диамантен* има релация *sem-derived-from* с *диамант*.

Тази релация би била ползотворна и за установяване на връзки между наречия и съществителни.

sem-derives-to

Релацията *образува* (*sem-derives-to*) се използва за съществителни, които са семантично или словообразователно свързани с прилагателни. Например: *дантела* има релация *sem-derives-to* с *дантелен*.

sem-derived-from-v

Релацията *образуван_от_глагол* (*sem-derived-from-v*) се използва за прилагателни, които са семантично или словообразователно свързани с глагол. Например: *административен* има релация *sem-derived-from-v* с *администрирам*.

sem-derives-to-v

Релацията *образува_глагол* (*sem-derives-to-v*) се използва за глаголи и отбелязва прилагателните, с които те са свързани. Например: *администрирам* има *sem-derives-to-v* релация с *административен*.

sem-derived-from-adj

Релацията *образуван_от_прилагателно* (*sem-derived-from-adj*) се използва за наречия, свързани с прилагателни имена. Например: *тъмно* има *sem-derived-from-adj* към *тъмен*.

sem-derives-to-adj

Релацията *образува_прилагателно* (*sem-derives-to-adj*) показва с кои наречия е свързано дадено прилагателно. Например: *артистичен* има *sem-derives-to-adj* релация към *артистично*.

predecessor

Релацията *предшественик* (*predecessor*) се използва само за именувани същности и свързва дадено понятие с неговия наследник. Например: *Университет за национално и международно стопанство* има релация *predecessor* към *Висш икономически институт „Карл Маркс“*.

successor

Релацията *наследник* (*successor*) се използва само за именувани същности и свързва

дадено понятие с неговия предшественик. Например: *Свободен университет за политически и стопански науки* има релация *successor* към *Държавно висше училище за финансови и административни науки*.

2. Между лемии

В ВTB-WN няма релации между лемии, основният фокус в текущата версия е върху релациите между синонимни гнезда. Въвеждането на релации между лемии е част от идеите за бъдеща работа по мрежата от думи. Ще има поне два вида такива релации - антонимия и словообразователна релация.

6. Съответствия

6.1. С Open English Wordnet (OEW)

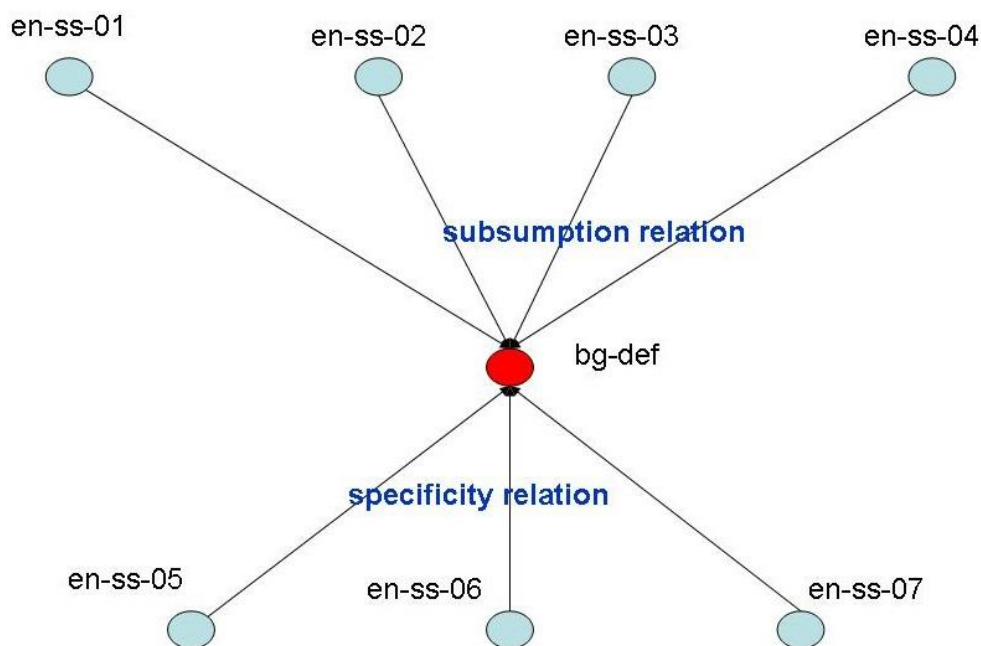
В този раздел е представено съответствието, което е установено с английската мрежа от думи OEW (първоначално с Princeton Wordnet 3.0).

Процесът на съпоставяне на значения в мрежата от думи е ориентиран към понятията - съпоставят се понятия, а не думи. Например българска дума може да е съпоставена с английска в съответствие с речник, но английската дума може да не отговаря на даденото понятие. В подобни случаи съпоставянето се прави със стратегии, различни от съпоставяне на думи (виж раздел 3).

ВTB-WN е съпоставен с Princeton WordNet 3.0, а по-късно и с OEW. Първо е направен ръчен превод на подгрупата синонимни гнезда от Core WordNet на PWN на български с цел да се постигне подходящо покритие между двата ресурса. Втората стъпка включва разпознаването на значения, открити в BulTreeBank, и организирането на тези значения в български синонимни гнезда, които след това са съотнесени към понятийната структура на PWN.

Процесът на съотнасяне започва с превод на български понятия на английски, след това търсене на съответстващо английско синонимно гнездо и установяване на релация между двете синонимни гнезда (релациите, използвани в ВTB-WN, са представени в раздел 5), добавяне на българска дефиниция и примери.

От 2020 г. ВTB-WN е съпоставен с OEW и основните ползи от това са, че тази мрежа от думи се обновява, редактира и разширява (за разлика от PWN). OEW е свързан с Уикиданни и Уикипедия (за английски), затова може да се използва за прехвърляне на знания към ВTB-WN.



6.2. Съответствие с Уикипедия

Два вида разширения на ВТВ-WN са направени - разширяване на наличните лемни с нови значения и разширяване с отделни представители.

За първата задача всички лемни в ВТВ-WN са сравнени със заглавията на статиите от българската Уикипедия и значенията от Уикипедия, които липсват в ВТВ-WN, са добавени в мрежата от думи с дефиниции и линкове към Уикипедия. Заглавията на съответните английски статии от Уикипедия са извлечени и използвани за подбор на правилно значение на английски и за подходящо синонимно гнездо от OEWS, с което да се направи съпоставяне.

Пояснителните страници в Уикипедия се оказват полезни, защото съдържат информация за свързани думи. За това съпоставяне са използвани три вида съответствия между ВТВ-WN и българската Уикипедия - равенство на значенията в двата ресурса, понятие, налично в Уикипедия, но липсващо в ВТВ-WN и именувана същност, представена в Уикипедия, но не и в мрежата от думи. В последните два случая са създадени нови синонимни гнезда и са установени релации с PWN.

6.3. Съответствие с DBpedia

Съпоставянето с DBpedia¹⁰ (свободно достъпна мрежа от знания с информация от Уикимедия) е използвано за втория тип разширяване на ВТВ-WN - с отделни представители. Именуваните същности в BulTreeBank са анотирани с унифицирани идентификатори на ресурс (URIs) от тази мрежа от знания, но понеже българската DBpedia е сравнително малка, българската Уикипедия също е използвана. Именуваните същности, които липсват в Уикипедия и DBpedia, са анотирани с класове от онтологията на DBpedia. По този начин отделни представители на класовете са автоматично

¹⁰ <https://www.dbpedia.org/>

класифицирани като отделни представители на съответстващите синонимни гнезда. Засега съпоставянето е направено с трите най-чести вида именувани същности в BulTreeBank - хора, места и организации - и е частично подкрепено и от съпоставянето с българската Уикипедия. Поради тази причина то е приложимо и към други мрежи от знания, произлизащи от Уикипедия.

7. Бъдеща работа

В BTB-WN се планира да бъдат направени разширения в няколко посоки. Първо, речниковото покритие на тезауруса постоянно ще бъде увеличавано. Нововъведенията ще бъдат не само от ежедневния език, но от различни специализирани области, от разговорната реч, диалектите и др. Значенията в BTB-WN ще бъдат съотнесени с различни онтологии. По този начин ще се осигури достъп до базите от знание. Друга задача, по която вече се работи, е съотнасянето на BTB-WN с валентен речник за български език и с Predicate Matrix¹¹ (лексикален ресурс, който интегрира FrameNet, VerbNet, PropBank, WordNet и ESO). Първата задача е част от нашия план за интегриране на всички български ресурси, а втората ще тества нивото на автоматично пренасяне на знание между английски и български с последващо редактиране. Задачата е релевантна, защото английският продължава да бъде един от езиците с повече и по-добре интегрирани езикови ресурси. Като първоначална проба са анотирани текстове от паралелните английско-български части на корпуса SETimes corpus¹² - първо, английската част от тях е анотирана със значения от Predicate Matrix и тези значения са прехвърлени в българската част.

Няколко случая са забелязани при прехвърлените значения: (1) българското значение се различава от това в Predicate Matrix, но все пак е валидно; (2) няколко значения съвпадат, включително правилното; (3) едно значение е съвпаднало от BTB-WN и то е правилното; (4) съотнесено е грешно значение, защото съответстващото значение на лемата липсва в BTB-WN. Има и значения, които не са се прехвърлили заради грешки в тагера за части на речта (английски или български) или заради липсваща лема в BTB-WN. Затова се предвиждат нови експерименти, които ще използват BTB-WN с по-добро покритие.

8. Източници

Тук е представен списък с основните източници, които са използвани по време на създаването на BulTreeBank WordNet за определяне на възможни значения, формулиране на дефиниции, примери, съотнасяне с Princeton WordNet и English Open WordNet.

¹¹ <https://adimen.si.ehu.es/web/PredicateMatrix>

¹² <http://nlp.ffzg.hr/resources/corpora/setimes/>

1. Многоетомен речник на българския език (<https://ibl.bas.bg/rbe/>)
2. Л. Андрейчин и др. Български тълковен речник, IV издание, допълнено и преработено от Д. Попов. Наука и изкуство, 1994
3. Е. Пернишка, Д. Благоева и С. Колковска. Речник на новите думи в българския език, Наука и изкуство, 2010
4. Д. Благоева и С. Колковска. Речник на новите думи в българския език, Наука и изкуство, 2021
5. И. Касабов и К. Стоянов. Универсален енциклопедичен речник, Свидас, 1999, 2003
6. А. Нанова. Български синонимен и антонимен речник с идиоми, Просвета, 2019
7. Онлайн речник (<http://www.onlinerechnik.com/>)
8. Princeton WordNet (<https://wordnet.princeton.edu/>)
9. English Open WordNet (<https://en-word.net/>)
10. Уикипедия (https://bg.wikipedia.org/wiki/Начална_страница)
11. Уикиречник (https://bg.wiktionary.org/wiki/Уикиречник:Начална_страница)

В процеса на създаване на BTB-WN са използвани и конкордансери на няколко български корпуси:

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

9. Библиография

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)