



**ЕВРОПЕЙСКИ СЪЮЗ**  
ЕВРОПЕЙСКИ ФОНД ЗА  
РЕГИОНАЛНО РАЗВИТИЕ



ОПЕРАТИВНА ПРОГРАМА  
**НАУКА И ОБРАЗОВАНИЕ ЗА  
ИНТЕЛИГЕНТЕН РАСТЕЖ**

# Стратегии за подобряване на покритието на българския уърднет ВТВ-WordNet

Зара Кънчева  
Изкуствен интелект и езикови технологии  
ИИКТ – БАН



ЦЕНТЪР ЗА ВЪРХОВИ ПОСТИЖЕНИЯ ПО  
ИНФОРМАТИКА И ИНФОРМАЦИОННИ И  
КОМУНИКАЦИОННИ ТЕХНОЛОГИИ



- ✓ Увод
- ✓ VulTreeBank-WordNet
- ✓ Подобряване на покритието на VTB-WN
- ✓ Заключение
- ✓ Използвана литература



- ✓ Научен проект № 4: Езикови технологии и технологии, базирани на съдържание, за приложения над големи данни
- ✓ Задача № 4.1: Езикови технологии и технологии, базирани върху съдържание, за машинен превод



## ✓ Цели и задачи:

- развиване на елементи от ресурса VTB-WordNet за съпоставяне към формални онтологии (представяне на знание с множества от категории и свойства и връзки между тях) и за използване на методи за автоматично генериране на данни за обучение на векторно представяне на езикови данни (*word embedding*)
- свързване на VTB-WordNet с различни онтологии и мрежи от знание (*knowledge graph* – мрежа от обекти от реалния свят – обекти, събития, ситуации или понятия с връзки помежду си)
- експерименти с трениране на векторно представяне на езикови данни
- проверка и редакция на съществуващите синонимни множества в VTB-WN и обогатяване с нови значения от сравнение с българската Уикипедия





- ВТВ-WN е лексикален ресурс с характеристики на тълковен и синонимен речник, обогатен с енциклопедично знание (Osenova and Simov, 2018)
- сливане на лингвистично и енциклопедично знание (BabelNet (семантична мрежа, свързана с Wikipedia и Princeton WordNet), FrameNet (семантичен ресурс, свързан с PWN), Uby (PWN, GermaNet, Wiktionary, Wikipedia, FrameNet, VerbNet), VerbNet (синтактично-семантична мрежа, свързана с PWN и FrameNet), обединяването на PWN и Wikipedia и на pWordNet и PWN)
- подходящ за редица задачи при обработката на естествен език: сваляне на многозначност, извличане на релации, извличане и парсиране на именовани същности и на многокомпонентни думи, машинен превод и други



- понятията са организирани в синонимни множества (synsets) – групи от значения, подредени в йерархия, с различни релации (еквивалентност, хиперонимия, хипонимия и други)
- повече от 26 000 синсета в BVB-WN
- части на речта – съществителни, прилагателни и числителни имена, наречия, глаголи
- ръчен превод на Core WordNet от Princeton WordNet (Fellbaum, 1998) – 5000 базови понятия
- обогатяване със значения от българската трибанка BulTreeBank, честотен списък, българските Уикиречник и Уикипедия
- по-ранно разширяване на BVB-WN с данни от Уикипедия – 15% увеличение (Simov et al., 2019)



Лема: река

Синонимно гнездо

Част на реч	Дефиниция	Категория	Подрежда	№	BTB id	EN id	Идентификатор	Заклучен
n	Голяма маса вода, която тече в естествено корито и се влива в море, езеро или в друга река.	noun.object	0	1	btbwn-017000267-n		140346	F
n	Голямо множество, огромен брой хора, животни, предмети или други, които се движат в една по	noun.group	0	2	btbwn-014000472-n		140888	F
v	Давам указания, нареждания на някого да направи, да изпълни нещо.	verb.communication	0	3	btbwn-032000003-v	ewn-0074870	16264	F
v	Имам определено значение, смисъл.	verb.communication	0	4	btbwn-032000330-v	ewn-0095710	16579	F

Лексикална единица

40 примера, 40 от които към лема

Лема	Пример	# Примери	Идентифика	Подрежда	Част на реч
река	Хотелът на маските е в центъра на курорт	40	191368	1	n

Концептуални релации

Надлонятия / подлонятия

Допълнителна информация

Проблеми / въпроси

Отворен въпрос

Времени бележки

Голяма маса вода, която тече в естествено корито и се влива в море, езеро или в друга река.

**instance\_hyponym**

Лема	Дефиниция	Част на реч	Еквивалент	Идентифика
Ахерон	В древногръцката митология е име на река, която тече в подземното царство.	n	E	11384
Амазонка	Река в северната част на Южна Америка.	n	E	11391
Амур	Голяма река в далечния изток на Русия, влива се в Охотско море.	n	E	11393
Колорадо	Река в югозападната част на Съединените американски щати и Северозападно Мексико.	n	E	11433
Дунав	Втората по дължина река в Европа, която извира от Германия и се влива в Черно море.	n	E	11443
Делтауеър	Река в източната част на САЩ.	n	E	11445
Днепър	Четвъртата по дължина река в Европа, протичаща по териториите на Русия, Беларус и Украйна и се влива в	n	E	11448
Елба	Река в Централна Европа.	n	E	11451
Ефрат	Река в Западна Азия, която извира на територията на Турция от Кавказ и се влива в Индийския океан.	n	E	11454
Ганг	Река в Северна Индия и Бангладеш, която извира от западната част на Хималаите и стига до Бенгалския за	n	E	11467
Лена	Най-пълноводната река в Североизточен Сибир, извира от Байкалският хребет и се влива в Море Лаптеви.	n	E	11491
Маас,Мьоз	Голяма река в Западна Европа, която извира от Франция и се влива в Северно море.	n	E	11503
Мисисипи	Река в Северна Америка.	n	E	11506
Мисури	Река в Съединените американски щати, която извира от Скалистите планини и се влива в Мисисипи малко	n	E	11507
Нил	Река в североизточната част на Африка, втората най-дълга река в света.	n	E	11515
Об	Голяма река в азиатската част на Русия, Западен Сибир.	n	E	11518
Ориноко	Река във Венецуела, третата по големина река в Южна Америка.	n	E	11523
Рейн	Една от най-важните и големи реки в Европа, която извира от швейцарските Алпи и се влива в Северно мор	n	E	11552
Сакраменто	Най-дългата река в щата Калифорния, САЩ с дължина от 615 километра.	n	E	11557
Шелда	Река, която извира в северна Франция, пресича Белгия и се влива в Северно море.	n	E	11560
Стикс	В древногръцката митология - една от петте реки, които протичат през царството на Хадес, реката на омраз	n	E	11587
Тенеси	Река в САЩ с дължина 1049 километра.	n	E	11596
Темза	Най-голямата река в Англия.	n	E	11597
Висла	Най-важната и дълга река в Полша.	n	E	11610
Волга	Река в европейската част на Русия, най-дългата в Европа, влива се в Каспийско море.	n	E	11612
Дълга река,Яндзъ,Яндзъ	Най-голямата река в Китай и най-дългата река в Азия, която извира от Тибетската планинска земя и се влив	n	E	11616
Енисей	Река в Сибир, Русия, една от най-дългите и пълноводни реки в света.	n	E	11618
Aare River,Aare,Aar	a river in north central Switzerland that runs northeast into the Rhine	n	E	104042
River Acheron,Acheron	(Greek mythology) a river in Hades across which the souls of the dead were carried by Charon	n	E	104046
River Adige,Adige	a river in northern Italy that flows southeast into the Adriatic Sea	n	E	104052
Aire River,Aire,River Aire	a river in northern England that flows southeast through West Yorkshire	n	E	104065
Alabama,Alabama River	a river in Alabama formed by the confluence of the Coosa and Tallapoosa Rivers near Montgomery; flows southw	n	E	104066

Legend for pie chart:

- equivalent-to (3110)
- has\_domain\_topic
- holo\_part
- hyponym
- hyponym
- instance\_hyponym (235)
- mero\_part
- sem-derives-to



- ✓ Уикипедия:
  - богат и постоянно разширяващ се източник на информация
  - 277 076 статии на български език
  - понятия от разнообразни сфери – специализирани термини и общоупотребима лексика
  - възможност за тематичен подход – обогатяване със значения от определена област, защото Уикипедия предлага разглеждане на понятия, разпределени в собствени категории
  - лесно наблюдение на йерархията на категориите и избор на нива, от които да се извлекат понятия
  - и двата ресурса съдържат понятия и отделни случаи на понятия – Уикипедия има голям обем от имена на хора, географски обекти и събития (спортни състезания, концерти, войни и прочие), които не са застъпени в ВТВ-WN







## Страници в категория „Военни звания“

Показани са 71 от общо 71 страници в тази категория.

- Военно звание
- \*
- Военни звания през Втората световна война
- Пагон
- Сигнифер
- A**
  - Ага
  - Адмирал
  - Адютант
  - Армейски генерал
  - Атаман
- Б**
  - Багатур
  - Бей
  - Бейлербей
  - Бригаден генерал
- В**
  - Велик войвода
  - Войвода
  - Войник
- Г**
  - Гардемарин
  - Генерал
  - Генерал от авиацията
  - Генерал от артилерията
  - Генерал от артилерията (Руска империя)
  - Генерал от кавалерията
- Генерал от свързочните войски
- Генерал-адмирал
- Генерал-аншеф
- Генерал-лейтенант
- Генерал-майор
- Генерал-полковник
- Генералисимус
- Главнокомандващ
- Гросадмирал
- Д**
  - Декурион (военен командир)
  - Джигит
- Е**
  - Ефрейтор
- К**
  - Канонир
  - Кадет
  - Капитан
  - Капудан паша
  - Комбриг
  - Кондотиер
  - Курсант
- Л**
  - Лейтенант
- М**
  - Майор
  - Мирмиран
  - Младши лейтенант
  - Младши сержант
- Офицерски кандидат
- П**
  - Паша
  - Подполковник
  - Полковник
  - Портупей-юнкер
  - Прапорщик
- Р**
  - Редник
- С**
  - Сержант
  - Старши лейтенант
  - Старши сержант
  - Старшина
  - Субалтерн-офицер
- Ф**
  - Фелдмаршал
  - Флигел-адютант
  - Флотилен адмирал
- Х**
  - Хауптман
- Ц**
  - Центурион
- Ч**
  - Чорбаджия
- Ш**
  - Шогун





- ✓ Ползи от разширяването на VTB-WN с понятия от Уикипедия:
- обогатяване на VTB-WN със синоними на съществуващи понятия, с нови понятия от разнообразни области и отделни случаи на понятия
- по-богати данни за векторно представяне
- по-качествена семантична анотация на текст



- P. Osenova, K. Simov, The Data-driven Bulgarian WordNet: BTBWN, Cognitive Studies | Études cognitives 18 (2018)
- C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998
- Rudnicka, E. K., Piasecki, M. T., Piotrowski, T., Łukasz Grabowski, and Bond, F. (2017). Mapping WordNets from the perspective of inter-lingual equivalence. Cognitive Studies — Etudes cognitives , 17(1373):1–17
- McCrae, J. P. (2018). Mapping WordNet Instances to Wikipedia. In Proceedings of Ninth Global WordNet Conference, pages 62–69. The Global WordNet Association
- Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). Uby – a large-scale unified lexical-semantic resource based on Imf. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 580–590, Avignon, France, April. Association for Computational Linguistics
- Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence, 193:217–250
- Kiril Simov, Petya Osenova, Laska Laskova, Ivajlo Radev, and Zara Kancheva. 2019. Aligning the Bulgarian BTBWordNet with the Bulgarian Wikipedia. In Proceedings of the 10th Global WordNet Conference, 290-297
- Collin Baker. 2008. FrameNet, present and future. In Jonathan Webster, Nancy Ide, and Alex Chengyu Fang, editors, The First International Conference on Global Interoperability for Language Resources, Hong Kong. City University, City University
- Karin Kipper-Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. Ph.D. thesis, University of Pennsylvania



# Благодаря за вниманието!

