Recent Developments in BulTreeBank-WordNet (BTB-WN)

Kiril Simov[*] and Petya Osenova[*^]

Artificial Intelligence and Language Technology, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences[*] and Sofia University “St. Kl. Ohridski”[^]

Plan of the Talk

- Introductory words
- Extending and Linking BTB-WN
- BTB-WN based applications
- Conclusions

Introduction

- The reported work here refers to the last three years (2020, 2021, 2022)
- It turned out that many NLP applications required **not only available resources but also appropriate integration among them**
- We started to view BTB-WN as a hub for linking grammar, other lexical data and world knowledge
- Our ultimate goal however would be that users could customize their own dictionaries, examples or other material through interlinked resources. In short: Maximum re-use of existing resources and contribution from different communities in building new ones!

Some History

- The development of BTB-WN goes back to the times when an *Ontology-based lexicon for Bulgarian* was initially constructed (**Simov and Osenova 2010**)
- Here we started with domain ontologies aligned to the upper ontology *DOLCE*, using *OntoWordNet* for introducing the middle level concepts
- The first version of BTB-WN was constructed by translation of Core WordNet and EuroWordNet Base concepts that were added to the Open Multilingual Wordnet

Some Broader Context

- The interface between lexical semantics and grammar, between lexicons and corpora has been extensively discussed from various points of view: linguistic, typological, formal, implementational, etc.
- We support the point of view in which the grammar is born in the lexicon, i.e. the *lexicalist-centric one*, without lowering the role of grammar at all. This is on a par with:
  - the linguistic theories that are constraint-based (such as **HPSG** and **LFG**) or are word-based (**dependency theories**)
  - the flagship project in **eLexicography - ELEXIS**

Rationale

- It is well-known that wordnets are thesauri. Despite providing the meanings of words grouped within synsets and relations among these synsets, they are still:
  - very static
  - self-contained and
  - often do not cover all parts of speech
- At the same time, they are good candidates for playing a central role – like a hub – for linking grammar, other lexical data and world knowledge

Rationale

- Thus, along with cleaning the meanings, relations, spellings within BTB-WN, we started also other tasks such as:
  - linking lemmas to their morphosyntactic characteristics (through a rich tagset and morphological/inflectional dictionary of Bulgarian)
  - linking meanings to examples from corpora
  - constructing a corpus of definitions, annotated with senses from BTB-WN
  - adding domain terms from SSH domains
  - adding dictionaries from previous times with their specific spellings
  - constructing a Bulgaria-centric knowledge graph as an extension of BTB-WN
  - aligning different ontologies with respect to BTB-WN
  - using the lexical chains over the BTB-WN graph for generating correct sense detection drills for learners of Bulgarian

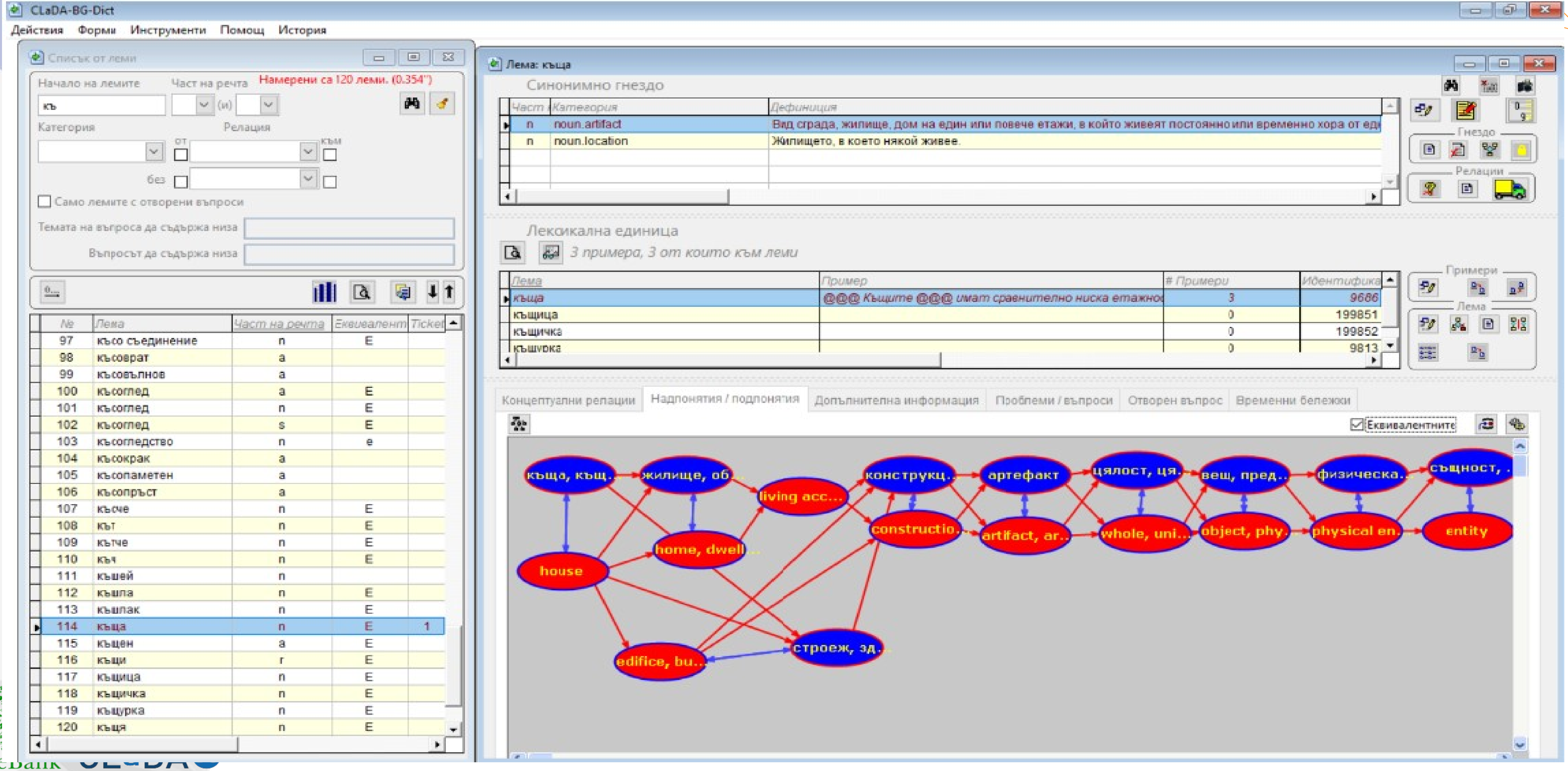
Linking of BTB-WN

- We switched from a tool that supported only local editing (where synsets were considered within a very limited context) (CLaRK system) to a tool that supports editing of the Wordnet data within a global context (CLaDA-BG-Dict)
- When a lemma is selected within BTB-WN, the following information can be accessed immediately:
  - the number of synsets related to it with the part-of-speech, as well as
  - the numbered meanings and links to the **Open English WordNet**
- The usage of almost each synonym within a synset is illustrated with examples
- Within the system the user could consult several other sources of information. The center of the system is BTB-WN

Linking of BTB-WN (2)

- The user could open as many editor forms as necessary in which to observe the synsets for different words
- The **Open English Wordnet** is available within the system
- The creation of a new Bulgarian synset could start from scratch entering all the information, including relations. But it is also possible to create such a synset with using an equivalent English synset
- In this way the relations of the English synsets are automatically transferred to BTB-WN

The Editor System CLaDA-BG-Dict



Списък от леми

Начало на лемите: къ

Част на речта (и)

Намерени са 120 леми. (0.354")

Категория

Релация

от □ към □

без □ □

□ Само лемите с отворени въпроси

Темата на въпроса да съдържа низ

Въпросът да съдържа низ

№	Лема	Част на речта	Еквивалент	Ticket
97	късо съединение	n	E	
98	късоврат	a		
99	късовълнов	a		
100	късоглед	a	E	
101	късоглед	n	E	
102	късоглед	s	E	
103	късогледство	n	e	
104	късокрак	a		
105	късопаметен	a		
106	късопръст	a		
107	късче	n	E	
108	кът	n	E	
109	кътче	n	E	
110	кърч	n	E	
111	кьшей	n		
112	къшла	n	E	
113	къшлак	n	E	
114	къща	n	E	1
115	къщен	a	E	
116	къщи	r	E	
117	къщица	n	E	
118	къщичка	n	E	
119	къщурка	n	E	
120	къщя	n	E	

Лема: къща

Синонимно гнездо

Част	Категория	Дефиниция
n	noun.artifact	Вид сграда, жилище, дом на един или повече етажи, в който живеят постоянно или временно хора от ед
n	noun.location	Жилището, в което някой живее.

Лексикална единица

3 примера, 3 от които към леми

Лема	Пример	# Примери	Идентифика
къща	@@@ Къщите @@@ имат сравнително ниска етажнос	3	9686
къщица		0	199851
къщичка		0	199852
къщурка		0	9813

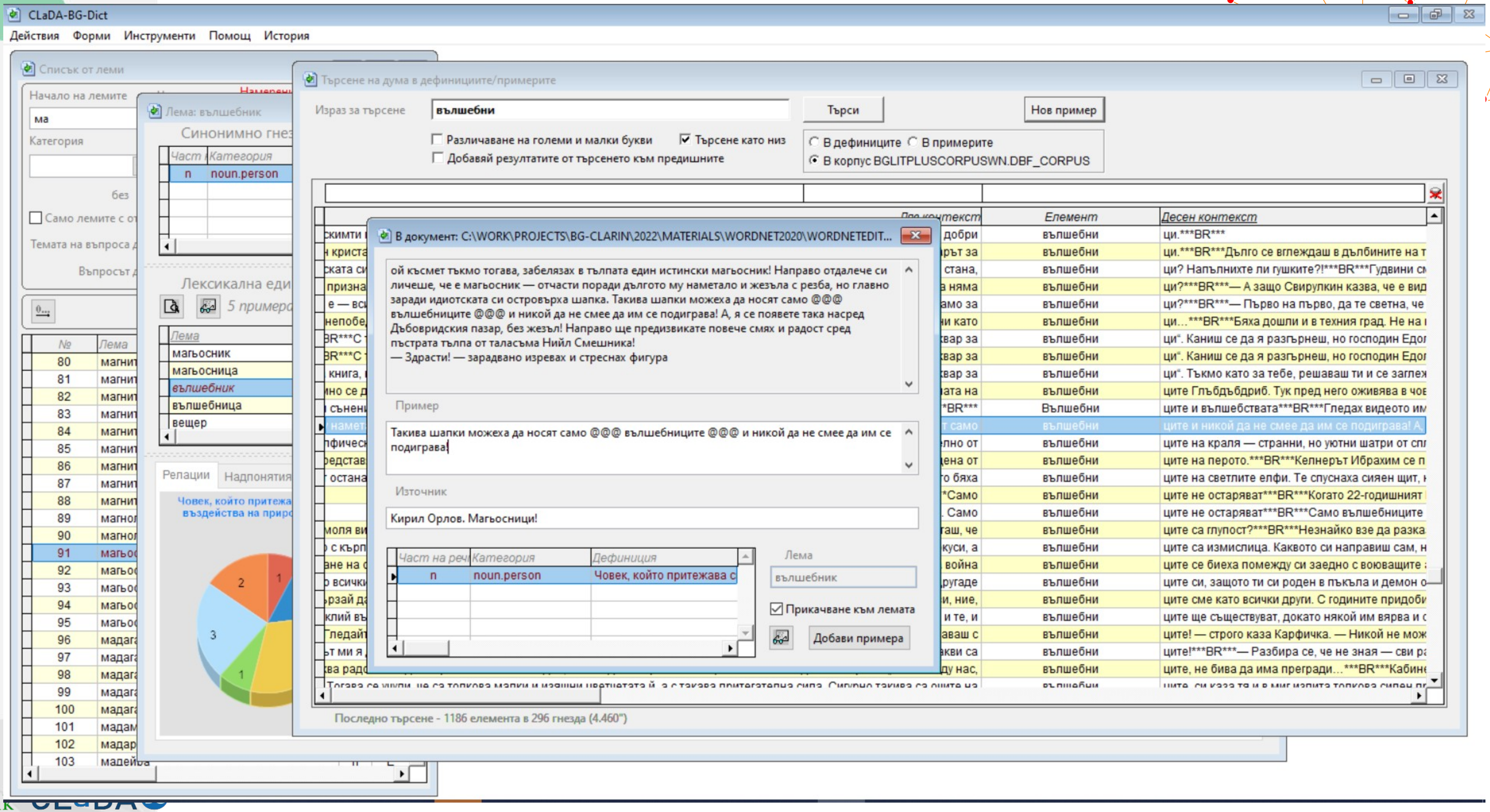
Linking of BTB-WN (3)

- In addition to granting access to OEW, the system provides access to dictionaries that are freely available to us, among which the **Bulgarian Explanatory Dictionary**, our **in-house Bulgarian Inflectional dictionary**, two **Bulgarian-English dictionaries**
- Each of these dictionaries could be consulted in isolation or simultaneously on the base of the alignments performed through lemmas
- The user could also define different lists of lemmas which to be mapped to the vocabulary of BTB-WN and to the vocabularies of the included dictionaries

Search in Corpora

- In addition to the data access options one can search with the selected lemma in *various text corpora*
- We consider the *definitions* and *examples* already included in BTB-WN as a corpus from which to select examples for other senses. In this case we could construct sense annotated corpora similar to *GlossCorpus*, *SemCor*
- The user could upload their own corpora when necessary

Search in Corpora


CLaDA-BG-Dict
Действия Форми Инструменти Помощ История
Списък от леми
Лема: вълшебник
Търсене на дума в дефинициите/примерите
Търси
Нов пример
Релации
Надпонятия
2
1
3
1

Начало на лемите: ма

Категория

без

□ Само лемите с от

Темата на въпроса д

Въпросът д

№	Лема
80	магнит
81	магнит
82	магнит
83	магнит
84	магнит
85	магнит
86	магнит
87	магнит
88	магнит
89	магнол
90	магнол
91	магьо
92	магьо
93	магьо
94	магьо
95	магьо
96	мадага
97	мадага
98	мадага
99	мадага
100	мадага
101	мадам
102	мадар
103	мадейра

Синонимно гнез

Част	Категория
n	noun.person

Лексикална еди

5 примера

Лема
магьосник
магьосница
вълшебник
вълшебница
вещер

Човек, който притежа
въздейства на приро

Израз за търсене: **вълшебни**

□ Различаване на големи и малки букви ☑ Търсене като низ
□ Добавяй резултатите от търсенето към предишните

○ В дефинициите ○ В примерите
● В корпус BGLITPLUSCORPUSWN.DBF_CORPUS

	контекст	Елемент	Десен контекст
скимти	добри	вълшебни	ци.***BR***
н криста	рът за	вълшебни	ци.***BR***Дълго се вглеждаш в дълбините на т
ската си	стана,	вълшебни	ци? Напълнихте ли гушките?!***BR***Гудвини см
призна	а няма	вълшебни	ци?***BR***— А защо Свирулкин казва, че е вид
е — вси	само за	вълшебни	ци?***BR***— Първо на първо, да те светна, че
непобе	ни като	вълшебни	ци…***BR***Бяха дошли и в техния град. Не на
BR***С т	квар за	вълшебни	ци". Каниш се да я разгърнеш, но господин Едог
BR***С т	квар за	вълшебни	ци". Каниш се да я разгърнеш, но господин Едог
книга,	квар за	вълшебни	ци". Тъкмо като за тебе, решаваш ти и се заглех
ино се д	ата на	вълшебни	ците Глъбдъбдриб. Тук пред него оживява в чов
сънени	*BR***	Вълшебни	ците и вълшебствата***BR***Гледах видеото им
намета	т само	вълшебни	ците и никой да не смее да им се подиграва! А,
лфическ	лно от	вълшебни	ците на краля — странни, но уютни шатри от спл
редстав	дена от	вълшебни	ците на перото.***BR***Келнерът Ибрахим се п
т остана	о бяха	вълшебни	ците на светлите елфи. Те спуснаха сияен щит,
	*Само	вълшебни	ците не остаряват***BR***Когато 22-годишният
	. Само	вълшебни	ците не остаряват***BR***Само вълшебниците
моля ви	гаш, че	вълшебни	ците са глупост?***BR***Незнайко взе да разка
с кърп	куси, а	вълшебни	ците са измислица. Каквото си направиш сам, н
ане на с	война	вълшебни	ците се биеха помежду си заедно с воюващите
всички	другаде	вълшебни	ците си, защото ти си роден в пъкъла и демон о
рзай да	и, ние,	вълшебни	ците сме като всички други. С годините придоби
клий въ	и те, и	вълшебни	ците ще съществуват, докато някой им вярва и
Гледайт	аваш с	вълшебни	ците! — строго каза Карфичка. — Никой не мож
ът ми я	акви са	вълшебни	ците!***BR***— Разбира се, че не зная — сви ра
ква радо	ду нас,	вълшебни	ците, не бива да има прегради…***BR***Кабине
Тогава се уили, че са толкова малки и изящни цветчетата й, а с такава притегателна сила. Сигурно такива са очите на		вълшебни	ците, си каза тя и в миг изпита толкова силен пр

Последно търсене - 1186 елемента в 296 гнезда (4.460")

В документ: C:\WORK\PROJECTS\BG-CLARIN\2022\MATERIALS\WORDNET2020\WORDNETEDIT...

ой късмет тъкмо тогава, забелязах в тълпата един истински магьосник! Направо отдалече си личеше, че е магьосник — отчасти поради дългото му наметало и жезъла с резба, но главно заради идиотската си островърха шапка. Такива шапки можеха да носят само @@@ вълшебниците @@@ и никой да не смее да им се подиграва! А, я се появете така насред Дъбовридския пазар, без жезъл! Направо ще предизвикате повече смях и радост сред пъстрата тълпа от таласъма Нийл Смешника!
— Здрасти! — зарадвано изревах и стреснах фигура

Пример

Такива шапки можеха да носят само @@@ вълшебниците @@@ и никой да не смее да им се подиграва!

Източник

Кирил Орлов. Магьосници!

Част на речи	Категория	Дефиниция
n	noun.person	Човек, който притежава с

Лема: вълшебник

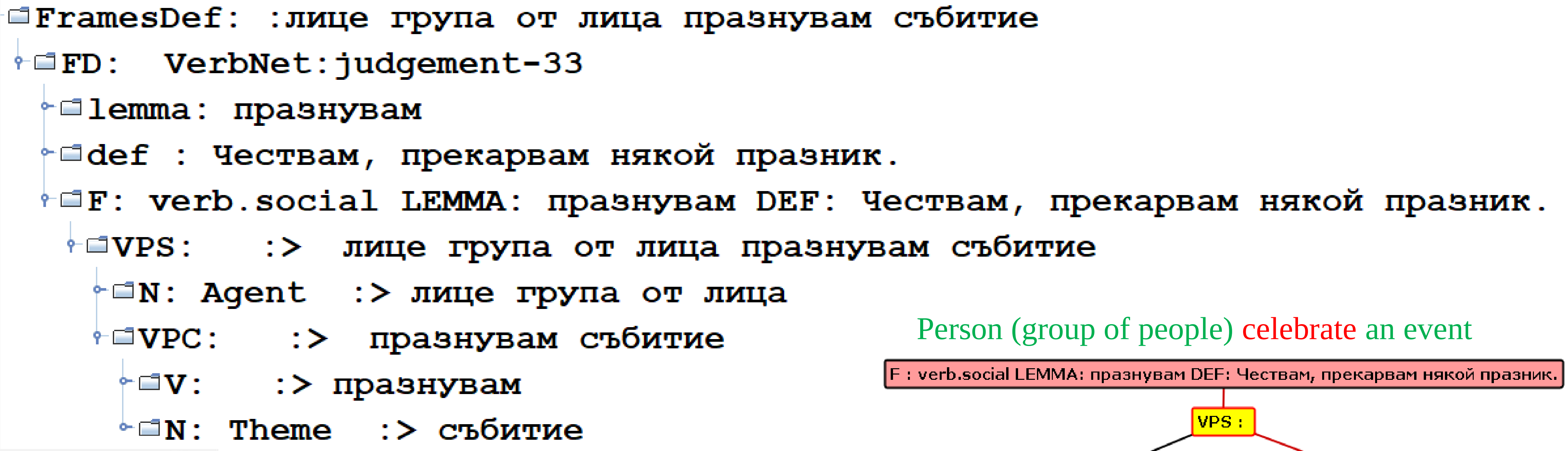
☑ Прикачване към лемата

Добави примера

Link to Valency Dictionary

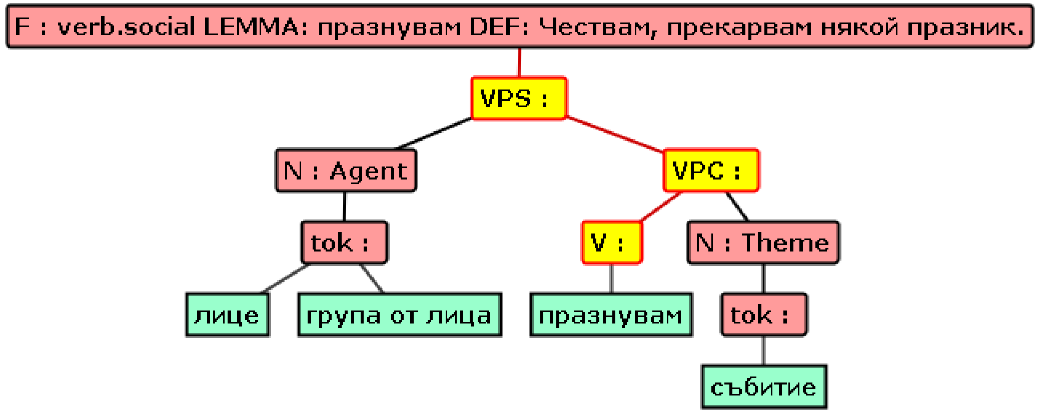
- The verb in a valency frame is connected to BTB-WN via a mapping to an appropriate synset from where access to the lexicographic class (such as *verb.social*, *verb.cognition*, etc.), the list of lemmas and the definition are available
- For example, if the *verb.emotion* worry is considered, the Bulgarian counterpart is displayed with a definition and a valency frame where the *Subject* has the role of *Experiencer* and the complement event that causes worrying has the role of *Stimulus*. The link to the **VerbNet frame** is also given

Example from the Valency Dictionary



```
FramesDef: :лице група от лица празнувам събитие
 FD:  VerbNet:judgement-33
  lemma: празнувам
  def : Чествам, прекарвам някой празник.
  F: verb.social LEMMA: празнувам DEF: Чествам, прекарвам някой празник.
   VPS:    :>  лице група от лица празнувам събитие
    N: Agent   :> лице група от лица
    VPC:    :>  празнувам събитие
     V:    :> празнувам
     N: Theme   :> събитие
```

Person (group of people) celebrate an event



F : verb.social LEMMA: празнувам DEF: Чествам, прекарвам някой празник.

VPS :

N : Agent

VPC :

tok :

V :

N : Theme

лице

група от лица

празнувам

tok :

събитие

From CLaRK to CLaDA-BG-Dict

When we switched from local processing in **CLaRK** system to the global processing in **CLaDA-BG-Dict** we had to perform examination of each synset in order to discover and repair every error that originated from the local processing. The synsets were checked with respect to the following criteria:

- Appropriateness of definitions
- Alignment to OEW
- Missing senses
- Wrong or missing relations
- Appropriateness of the examples

Appropriateness of Definitions

- We checked the definitions for the different kinds of word classes per synset
- This step was necessary, because we wanted to extend the definition to include more information than the definition within paper dictionaries
- This holds especially for adjectives. In the traditional dictionaries the adjective is usually defined as qualifying a noun. In our case we go further and develop the definition of the adjective also to the specific features of the qualified noun

Alignment to OEW

- In the previous versions having only a local view, we supported as many relations as possible between the Bulgarian and the English synsets some of which allowed in the noun and verb hierarchies to have disconnected elements
- With the switch to the global view it became much more convenient to verify these mappings and to re-consider some of them
- Now we focused on: *equivalent-to*, *hypernymy*, *homonymy* and *near-equivalent-to*

Missing Senses

- The construction of the early versions of BTB-WN were mainly driven by specific NLP tasks like Word Sense Disambiguation, Machine Translation, Mapping to Domain Ontologies
- In this applications we had to cover certain domains or type of texts. This resulted in representation of the senses of the different lemmas only partially
- Thus we decided to check the coverage of the resource with respect to the most common and well-established senses using the dictionaries available in the system (mainly)

Wrong or Missing Relations

- From the beginning we supported mapping to PWN (and now EOW). We are using this mapping to transfer automatically relations from English synsets to Bulgarian ones

- After the transfer the set of relations became eligible for modifications, if needed. This happens mainly when the mapping is not between equivalent synsets

Appropriateness of Examples

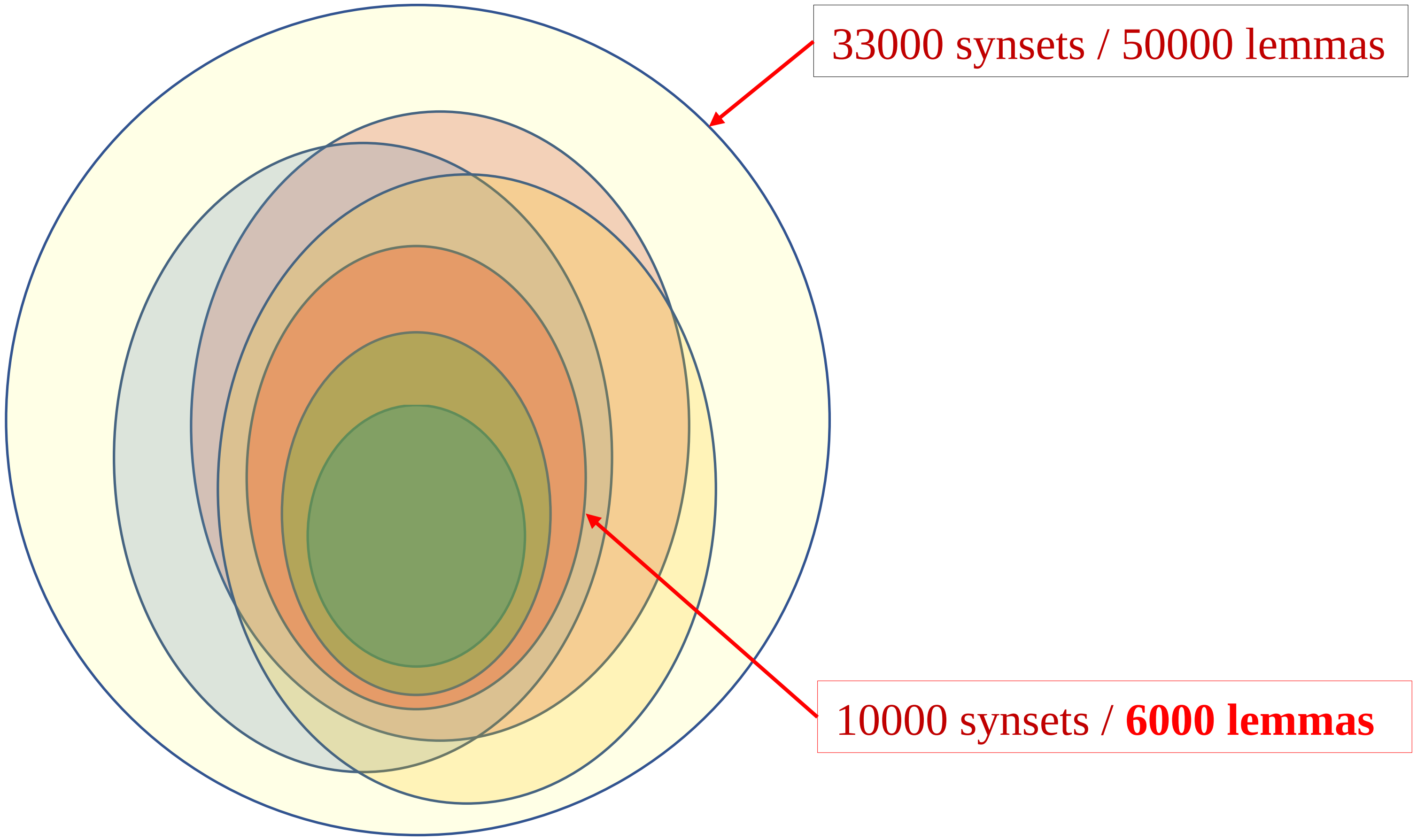
- The assigned examples were specially checked with respect to their appropriateness to the corresponding sense
- The most frequent error was when the example did not provide enough context for the meaning, and thus the corresponding word form might be interpreted ambiguously
- In such cases the example was extended or deleted
- Also we pay attention the examples to demonstrate as much as possible sense specific characteristics

Extensions

- Besides the examination of the existing synsets we extended BTB-WN with new synsets through the above mentioned vocabularies extracted from both types of sources - dictionaries and corpora
- Then the following information was added: derivational sets for these lemmas such as adjectives derived from nouns, aspectual variants of Bulgarian verbs that share a common basic sense, etc.
- In this way, more than **14000 synsets** were added (in total **33 000**)
- At the moment we completed the coverage of the core vocabulary with about **6000** lemmas.

Vocabulary Structure

- For different applications not everything is necessary
- In some cases additional requirements could be imposed like style and language of definitions, content of examples
- Some tasks are performed on the basis of the sub-vocabularies



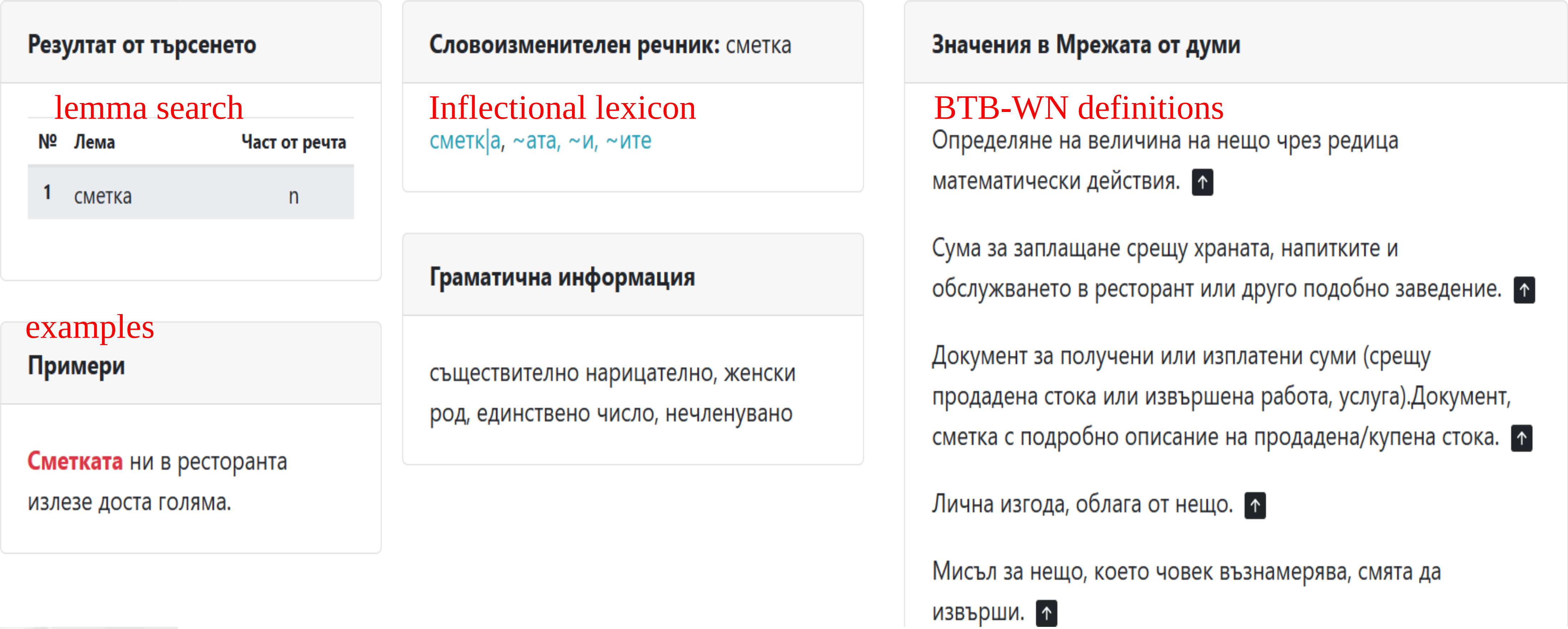
33000 synsets / 50000 lemmas

10000 synsets / **6000 lemmas**

BTB-WN based Applications (not NLP)

- **The Bulgaria-centric knowledge graph**
  - BTB-WN has been further enriched with terms from various Social Sciences and Humanities domains such as history and ethnography. Here two challenges appeared. The first one is related to the introduction of terminological multiword expressions while the second one refers to the register of usage such as being archaic or dialectal, etc.
- **The bigger net of dictionaries and resources, called in our case *All about words***
  - The system includes a concordancer, a Wordnet viewer, a word form analyser, a viewer for the Bulgarian inflectional dictionary, viewers for other dictionaries.

Integrated view



Резултат от търсенето

lemma search

№	Лема	Част от речта
1	сметка	n

examples

Примери

Сметката ни в ресторанта излезе доста голяма.

Словоизменителен речник: сметка

Inflectional lexicon

сметк|а, ~ата, ~и, ~ите

Граматична информация

съществително нарицателно, женски род, единствено число, нечленувано

Значения в Мрежата от думи

BTB-WN definitions

Определяне на величина на нещо чрез редица математически действия. ↑

Сума за заплащане срещу храната, напитките и обслужването в ресторант или друго подобно заведение. ↑

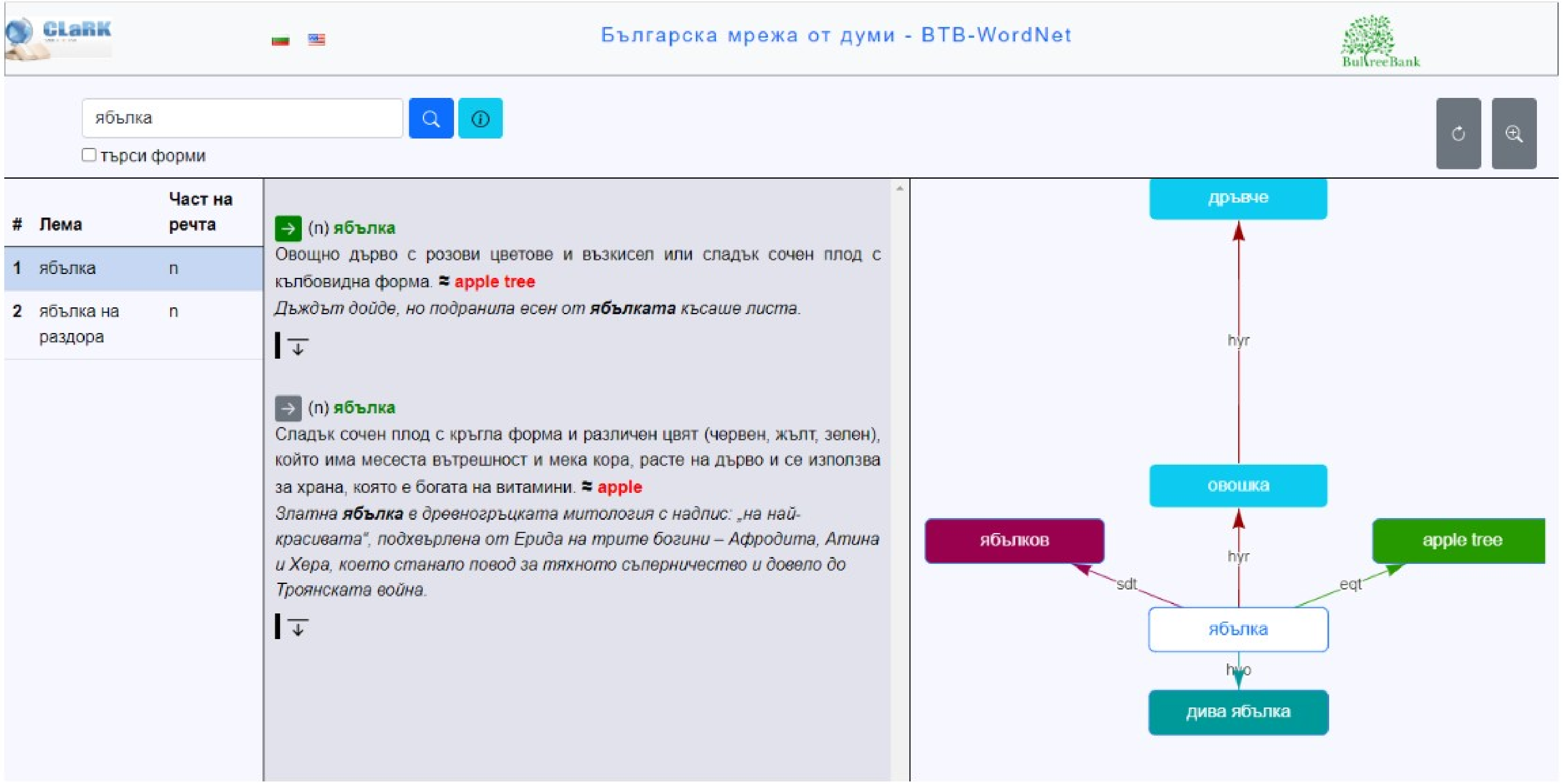
Документ за получени или изплатени суми (срещу продадена стока или извършена работа, услуга).Документ, сметка с подробно описание на продадена/купена стока. ↑

Лична изгода, облага от нещо. ↑

Мисъл за нещо, което човек възнамерява, смята да извърши. ↑

Missing part from the user interface: relations to other dictionaries

BTB-Wordnet Viewer



Българска мрежа от думи - BTB-WordNet

яблъка

□ търси форми

#	Лема	Част на речта
1	ябълка	n
2	ябълка на раздора	n

(n) **ябълка**

Овощно дърво с розови цветове и възкисел или сладък сочен плод с кълбовидна форма. ≈ **apple tree**

Дъждът дойде, но подранила есен от ***ябълката*** *късаше листа.*

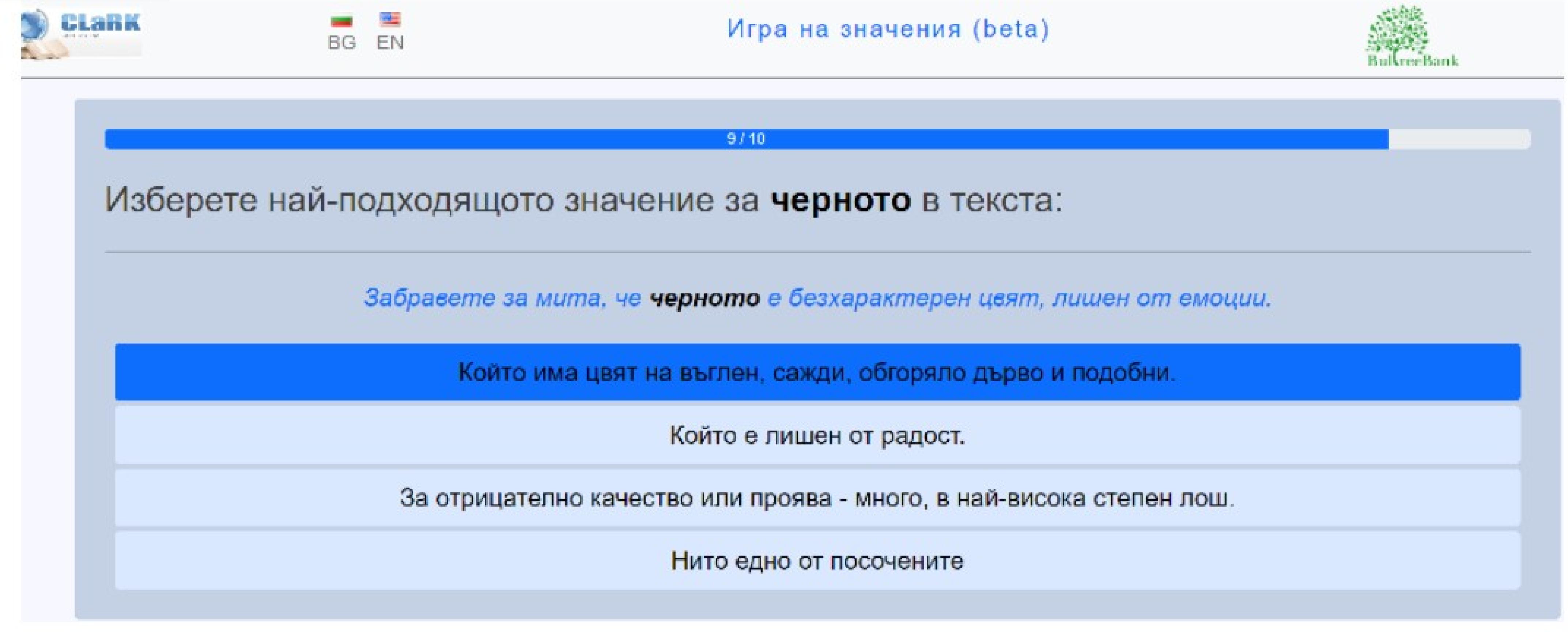
(n) **ябълка**

Сладък сочен плод с кръгла форма и различен цвят (червен, жълт, зелен), който има месеста вътрешност и мека кора, расте на дърво и се използва за храна, която е богата на витамини. ≈ **apple**

Златна ***ябълка*** *в древногръцката митология с надпис: „на най-красивата", подхвърлена от Ерида на трите богини – Афродита, Атина и Хера, което станало повод за тяхното съперничество и довело до Троянската война.*

Game of Meanings

Select the correct definition for the highlighted word in the example!



Игра на значения (beta)

Изберете най-подходящото значение за **черното** в текста:

Забравете за мита, че ***черното*** *е безхарактерен цвят, лишен от емоции.*

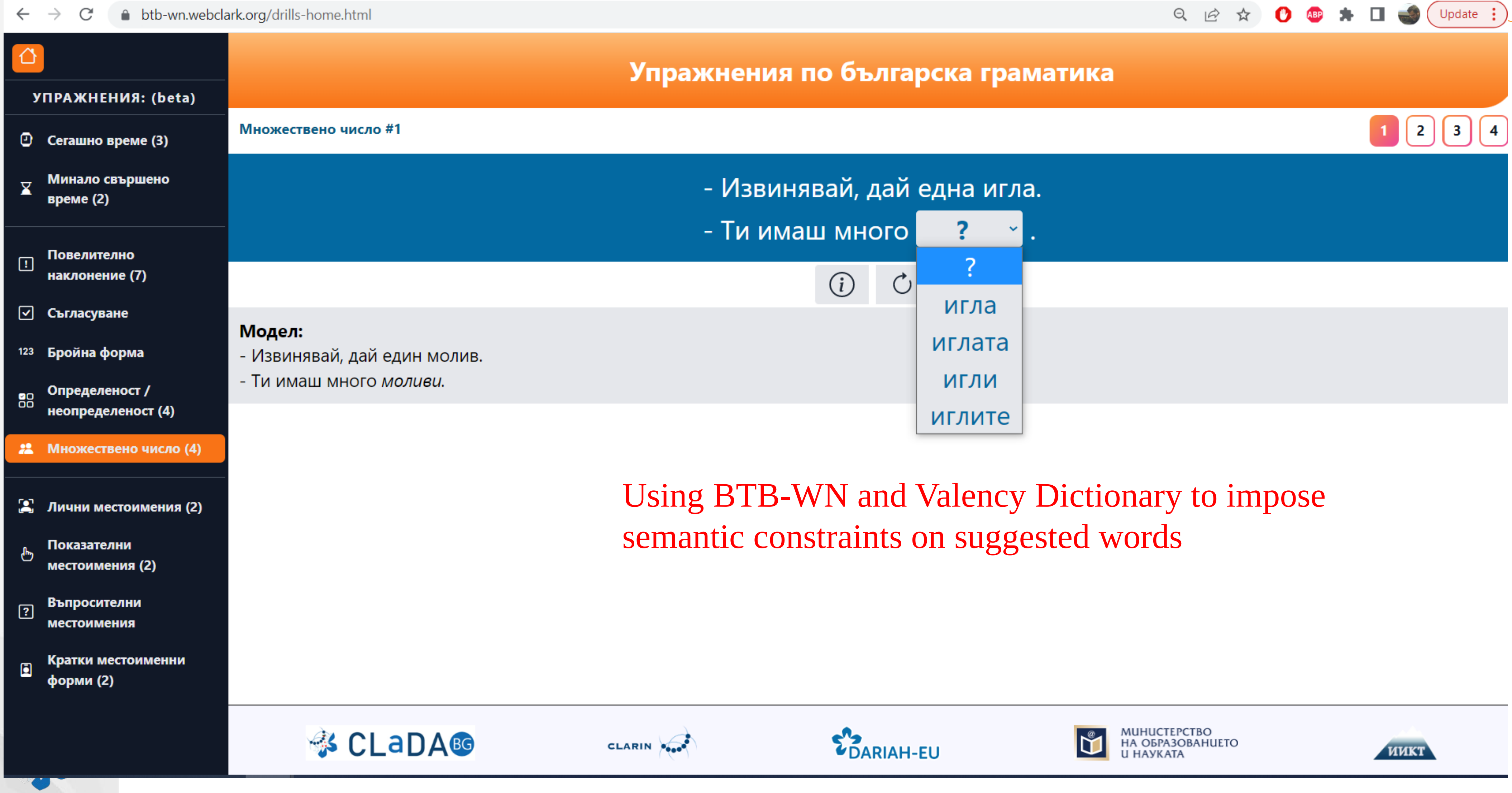
Който има цвят на въглен, сажди, обгоряло дърво и подобни.

Който е лишен от радост.

За отрицателно качество или проява - много, в най-висока степен лош.

Нито едно от посочените

Grammar drills


btb-wn.webclark.org/drills-home.html
Update
УПРАЖНЕНИЯ: (beta)
Сегашно време (3)
Минало свършено време (2)
Повелително наклонение (7)
Съгласуване
Бройна форма
Определеност / неопределеност (4)
Множествено число (4)
Лични местоимения (2)
Показателни местоимения (2)
Въпросителни местоимения
Кратки местоименни форми (2)
1
2
3
4
CLaDA BG
CLARIN
DARIAH-EU
МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА
ИИКТ

Упражнения по българска граматика

Множествено число #1

- Извинявай, дай една игла.

- Ти имаш много ? .

?
игла
иглата
игли
иглите

Модел:

- Извинявай, дай един молив.

- Ти имаш много *моливи.*

Using BTB-WN and Valency Dictionary to impose semantic constraints on suggested words

Conclusions

- We present the current 4.0 version of the BTB-WN with the following main new features (comparing to the previous versions):
  - more synsets with a better coverage on basic vocabularies
  - new types of definitions and examples
  - mapping to other resources: inflectional lexicon of Bulgarian (grammatical information and full paradigms of lemmas), Bulgarian Wikipedia
  - soon: mapping to Bulgarian Knowledge Graph and other lexical resources
- Implementation of NLP tasks
- Training of Language Models