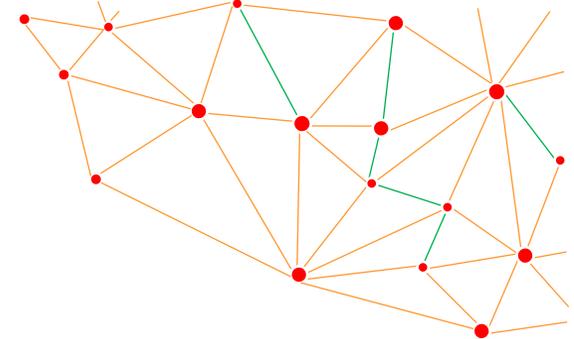


BTB-WordNet: Status and Challenges

Petya Osenova

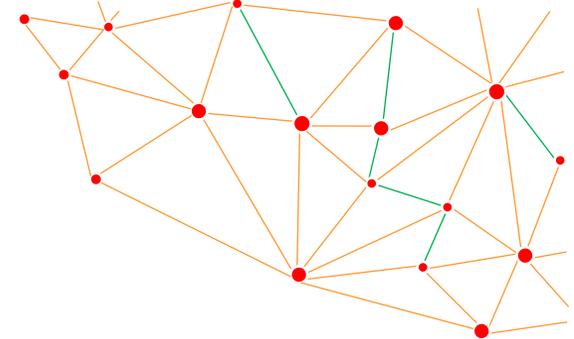
IICT-BAS and Sofia University “St. Kl. Ohridski”

Plan of the Talk



- Introduction
- Background of BTB-WordNet
- Challenges in mapping BTB-WordNet to the English wordnet
- Challenges in mapping BTB-WordNet to other resources
- Lessons Learnt

What a WordNet is?



- WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.
- WordNet resembles a thesaurus, in that it groups words together based on their meanings.

<https://wordnet.princeton.edu/>

Open English WordNet



Open English WordNet

LEMMA

university

OPTIONS ▼



Show Synset Identifier



Show Sense Keys



Show Subcategorization Frames



Show Topics



Show Pronunciation

Nouns

(n) university. *the body of faculty and students at a university*

MORE ►

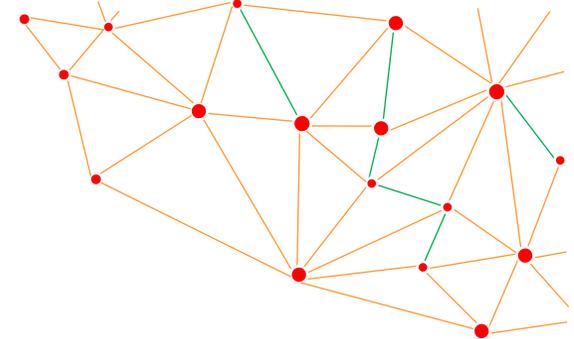
(n) university. *establishment where a seat of higher learning is housed, including administrative and living quarters as well as facilities for research and teaching*

MORE ►

(n) university. *a large and diverse institution of higher learning created to educate for life and for a profession and to grant degrees*

19.09.2023, Faculty of Linguistics Seminar, Iran

Open English WordNet



(n) university *a large and diverse institution of higher learning created to educate for life and for a profession and to grant degrees*

Hypernyms (1)

(n) educational institution *an institution dedicated to education*

Hypernyms (1)

(n) institution, establishment *an organization founded and united for a specific purpose*

MORE ►

Hyponyms (5)

(n) preschool *an educational institution for children too young for elementary school*

MORE ►

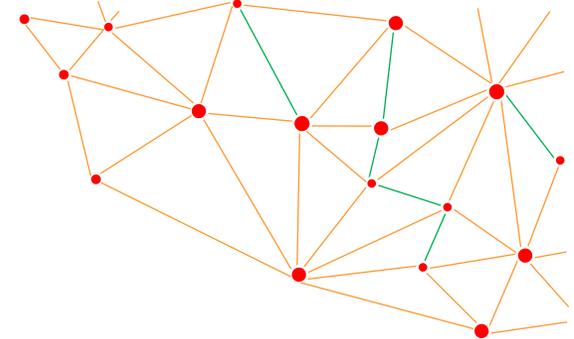
(n) school *an educational institution "the school was founded in 1900"*

MORE ►

(n) school *an educational institution's faculty and students "the school keeps parents informed" "the whole school turned out for the game"*

MORE ►

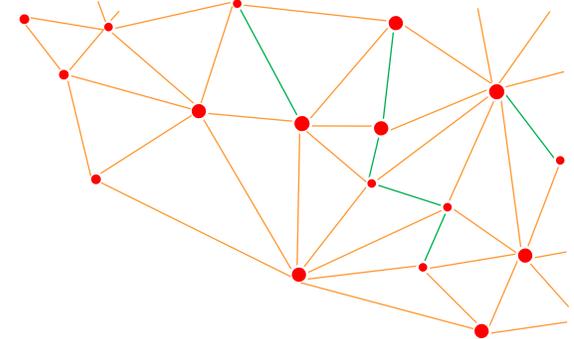
Introduction



Bulgarian BulTreeBank WordNet (BTB-WN) is created in three different stages:

- Translation of English synsets from Core WordNet subset of Princeton WordNet into Bulgarian
- Identification of senses used in Bulgarian Treebank BulTreeBank
- Sense extension: a) detection of missing senses of processed lemmas and adding them to BTB-WN; b) extraction of information from the Bulgarian Wiktionary
- Result – 19200 synsets (2019)

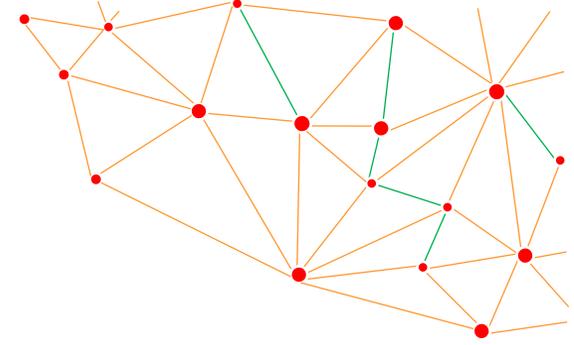
Introduction (2)



Recent development of BTB-WN:

- Currently, there are around 35 000 synsets and 49313 lemmas
- Editing the BTB-WN with a special software (started in the CLaRK system which is an XML system for language resources development, and now moved to BTBDict one)
- Under constant manual check are: a) the definitions of the synsets; b) the synonyms of each synset; c) the mapping to the EWN; d) examples

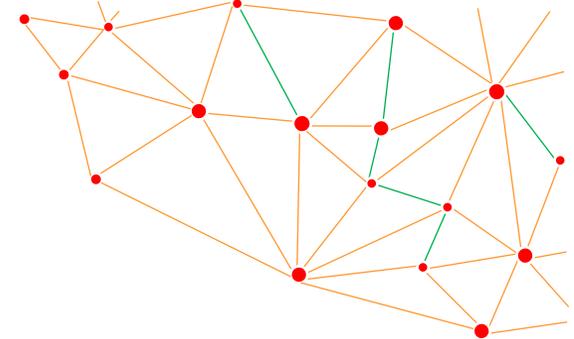
Introduction (3)



Problems with the mapping between languages in general:

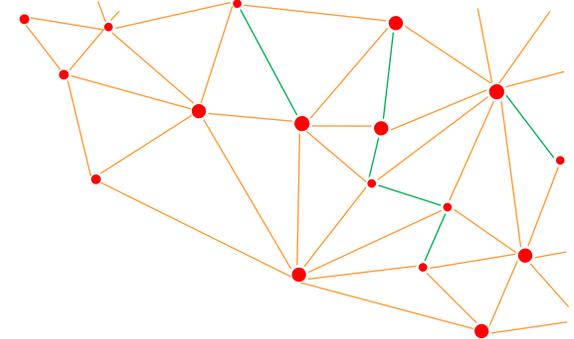
- WNs are not perfect: missing meanings, wrong classifications, etc.
- Languages differ in their linguistic properties
- WNs encode the cultural conceptualization reflected within the lexical bases of the languages

The Main Challenge



- The main challenge is the simultaneous creation of a data-driven WordNet for Bulgarian and a manually annotated treebank with semantic information
- It requires synchronization of the word senses in both - syntactic and lexical resources, without limiting the WordNet senses to the corpus or vice versa.

Parallel Semantic Enhancements

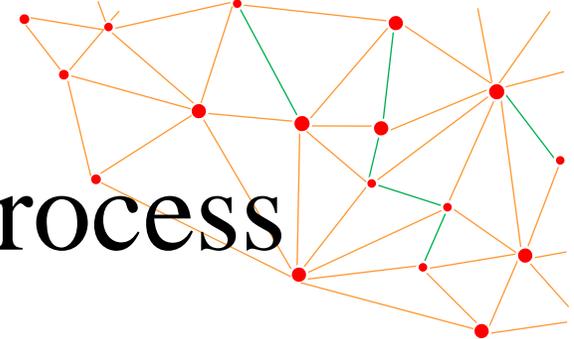


Semantic annotation of the treebank with Bulgarian resources



Extending the domain WordNets with general lexica

Our Approach Reflects the Creation Process

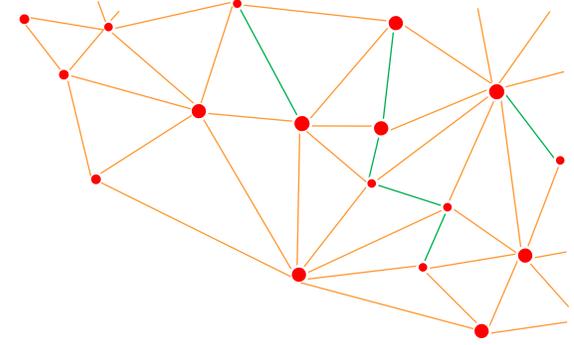


- *the expand method*: the translation of the synsets from the source into the target language
- *the merge method*: takes (also) into account the language specific resources

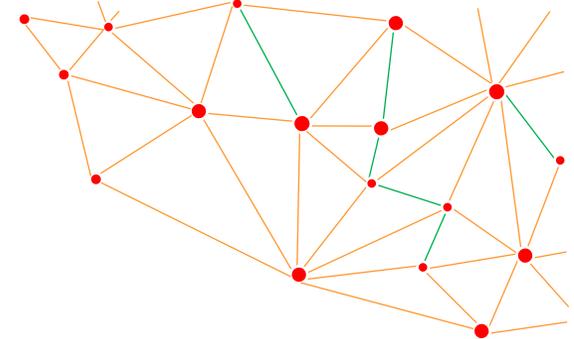
What combination of steps is the best? The one extreme, the other extreme, or some strategies in between?

Steps

- translation of English PWN into another language;
- data-driven approaches via identification of synsets within real texts;
- automatic extraction from existing lexical resources;
- various combinations of these.
- Also: automatic vs. manual work

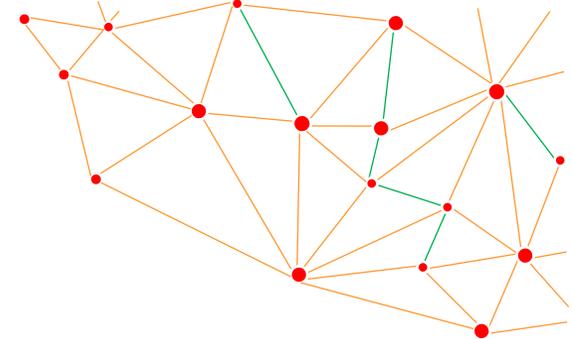


Sum up



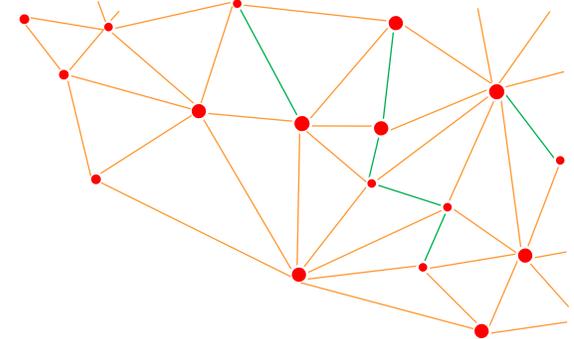
- There is no easy way to achieve typological consistency in building WordNets:
 - if the *expand* method is chosen, the language resource suffers from lack of nativeness of the hierarchy and relations;
 - if the *merge* method is followed, the language resource differs too much from other similar resources and it is time-consuming to map it back to them.

From the Corpora Perspective



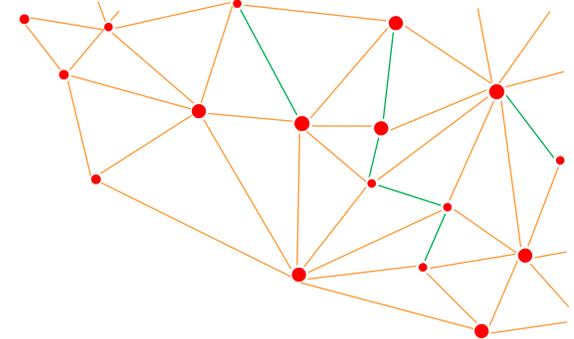
- The usual way of annotating senses in treebanks is the following:
 - there is a WordNet for the language in question, and then the treebank is annotated with senses from it
 - **HOWEVER**, they bear also the restrictions that are presented in the so-called static lexical resources
 - Therefore: the sparseness of the sense coverage might be really problematic.

Our Strategy



- **In-language sense annotation:** We first annotated the treebank with senses from an explanatory dictionary of Bulgarian;
- **Synset compilation:** Only then we started the formation of synsets;
- **Mapping:** They were then mapped to the PWN while keeping track of the various sense discrepancies by different mappings.

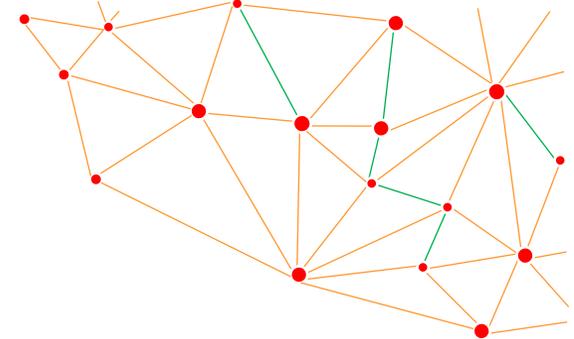
Coverage Problems



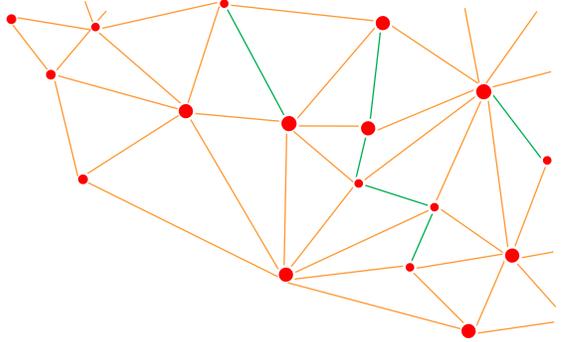
- The extensions on the basis of:
 - text annotation
 - the existing lexicon
- Exhibit the sparseness problem:
 - not all synonyms appear in the annotated texts and the lexical entries.

For that reason, we performed checks on the completeness of the synsets with respect to the missing synonyms.

Mid-Observations

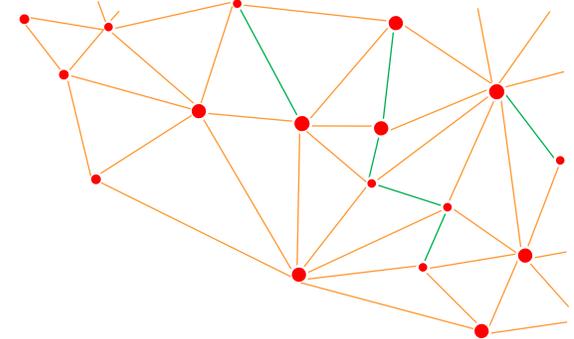


- We started from the **domain** semantic annotation.
- We were aware of the concept sparseness problem.
- We did not rely on pre-created WordNet, but rather on an explanatory dictionary of Bulgarian.
- Later mappings to PWN
- **Early involvement in various NLP applications**

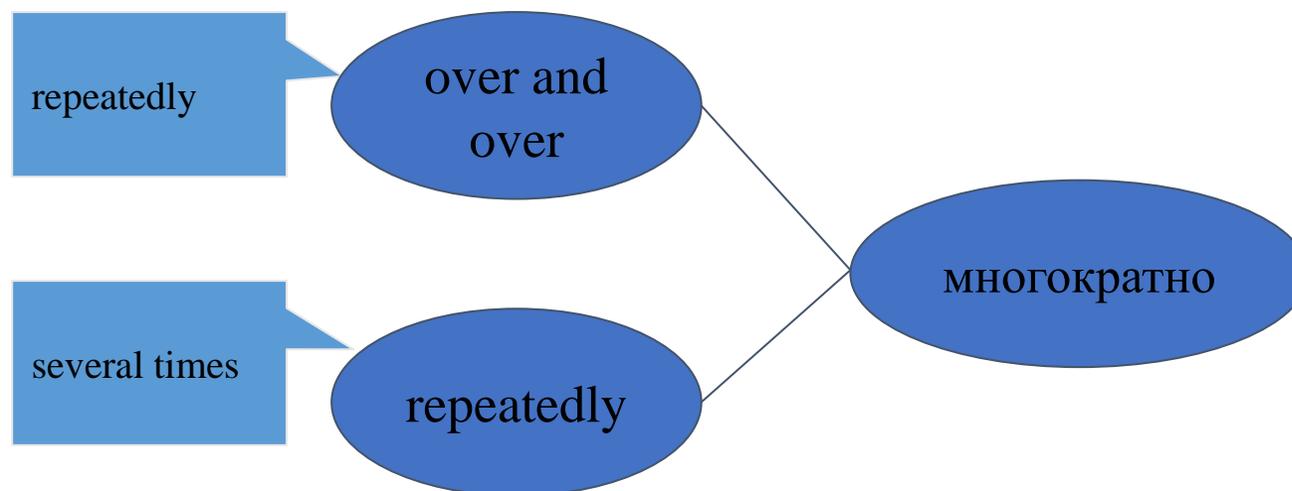


Delving into the Mapping to Princeton and English Wordnet

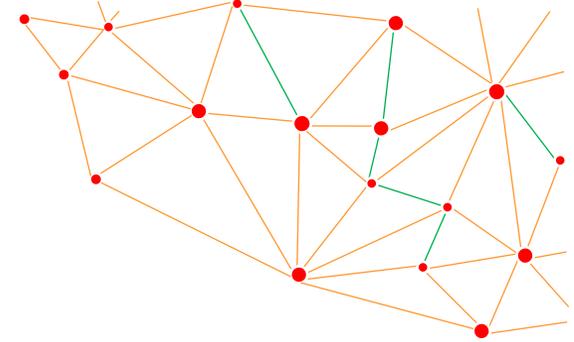
Problems of EWN Hierarchy



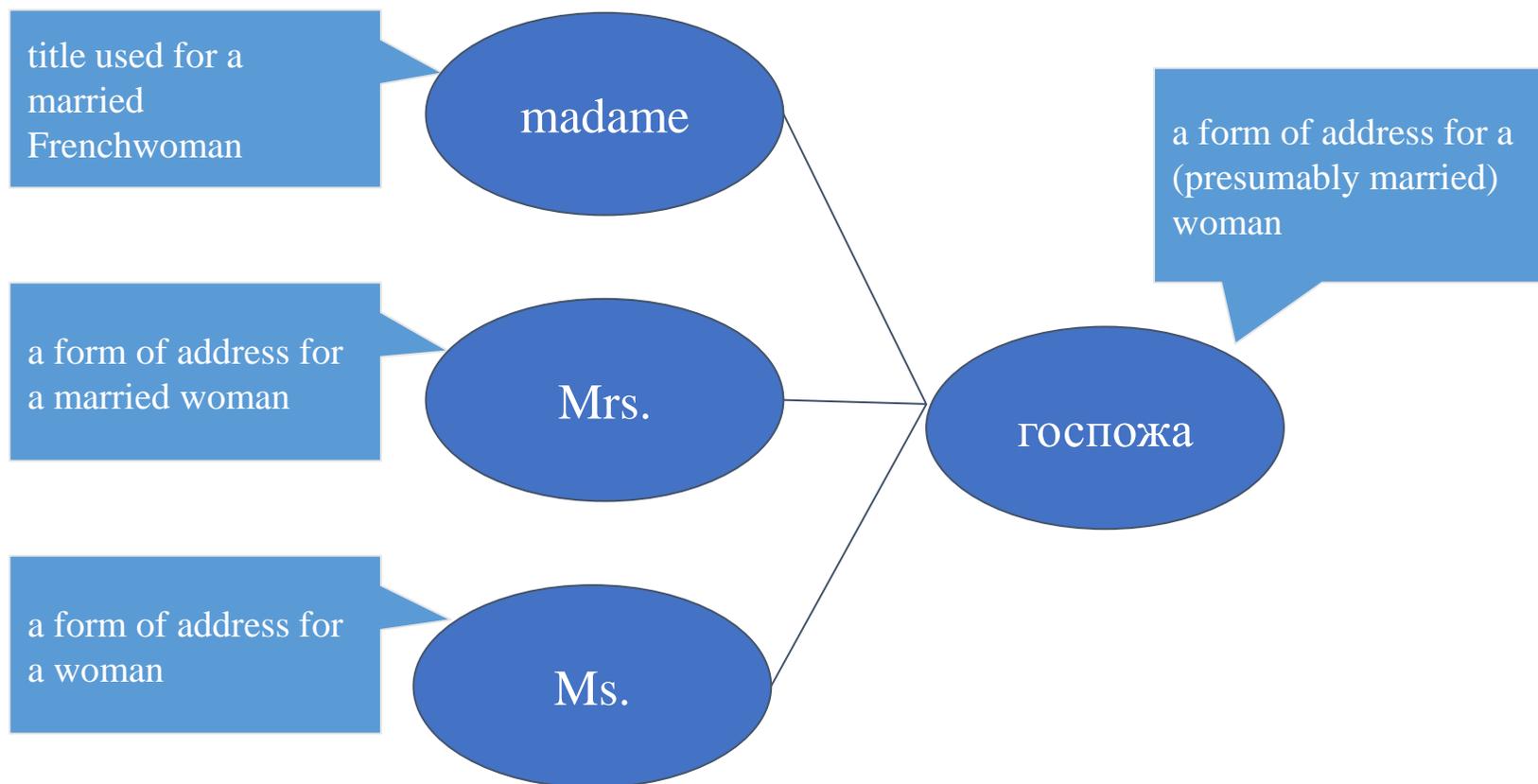
Differentiation with very subtle differences:



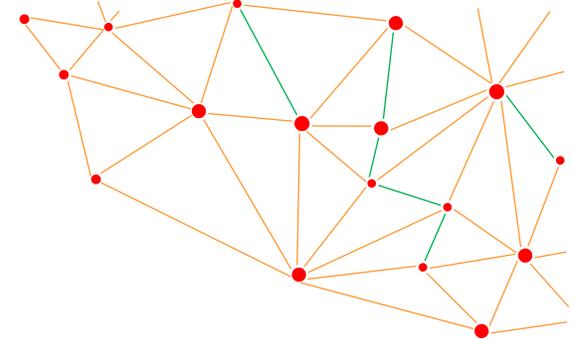
Problems of EWN Hierarchy



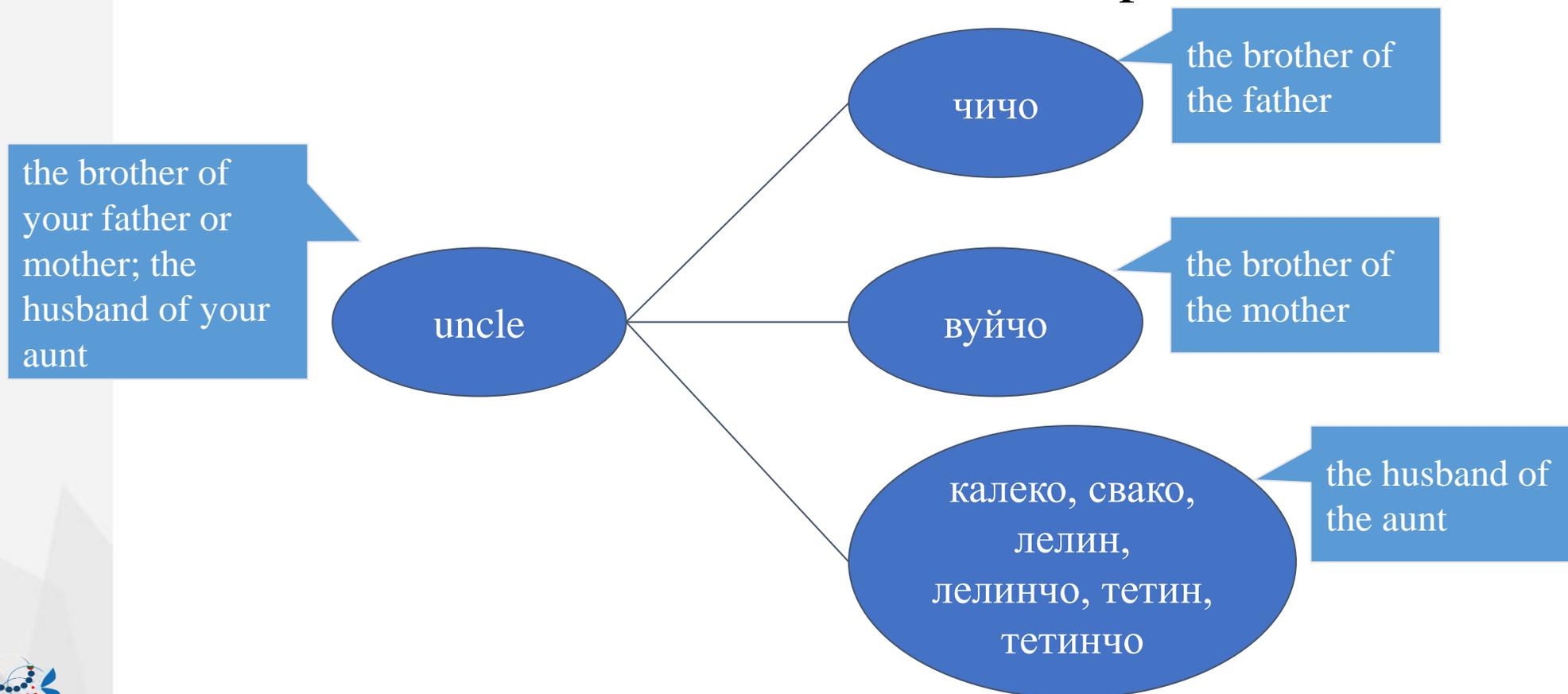
Differentiation with very subtle differences:



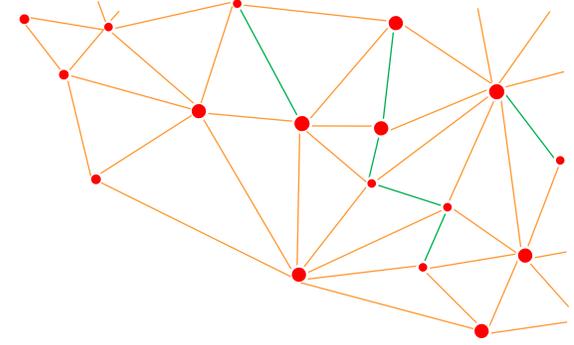
Problems of EWN Hierarchy



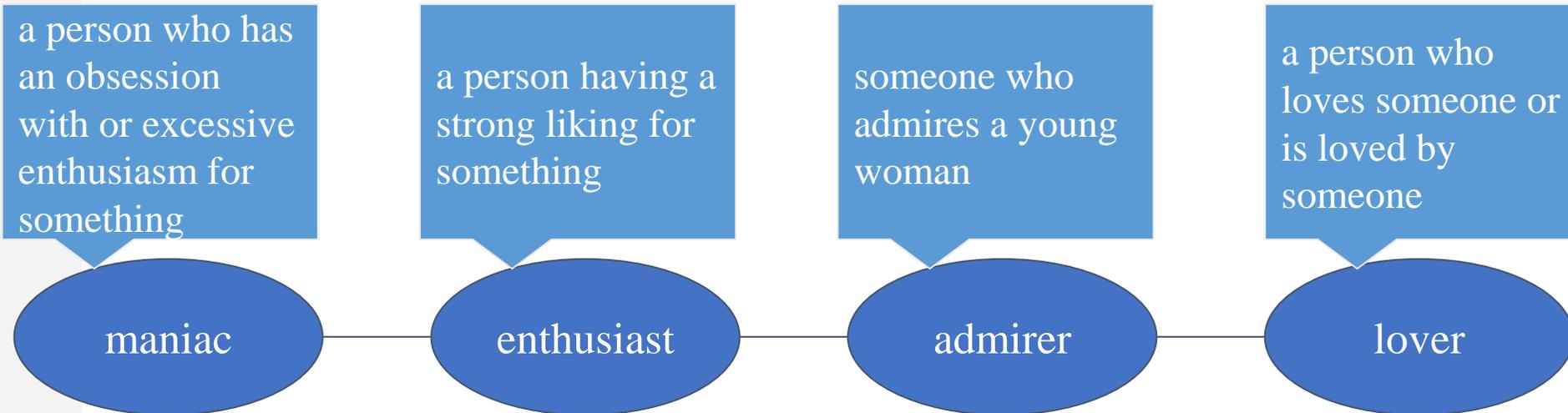
In some cases BTB-WN differentiates more concepts than EWN:



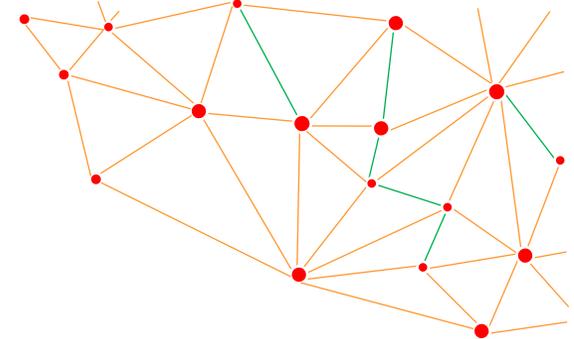
Problems of EWN Hierarchy



Changing the hierarchies as they are incorrect:

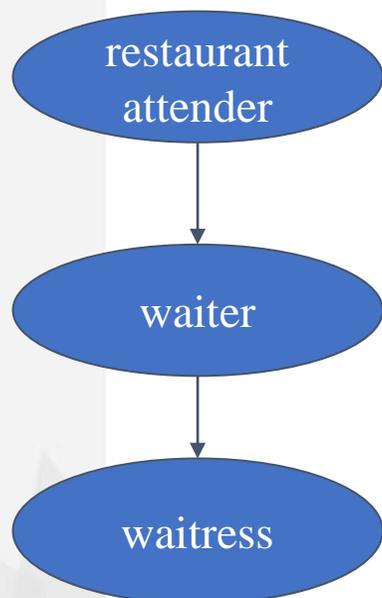


Problems of EWN Hierarchy

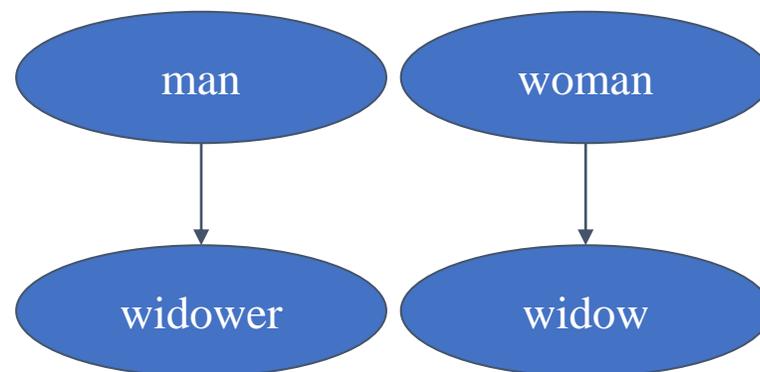


Male and female word forms:

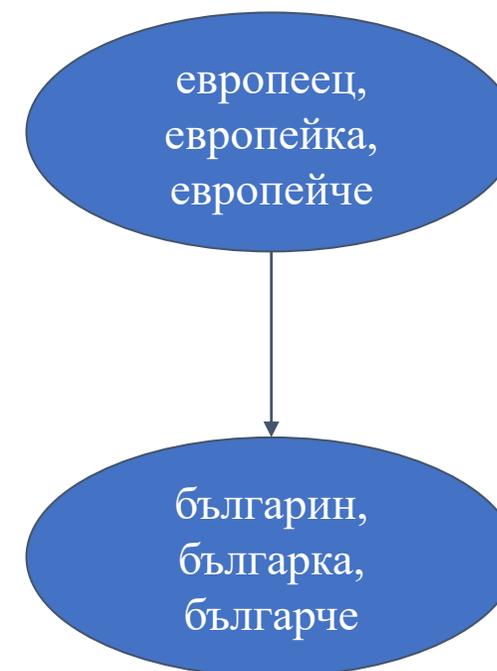
1. EWN



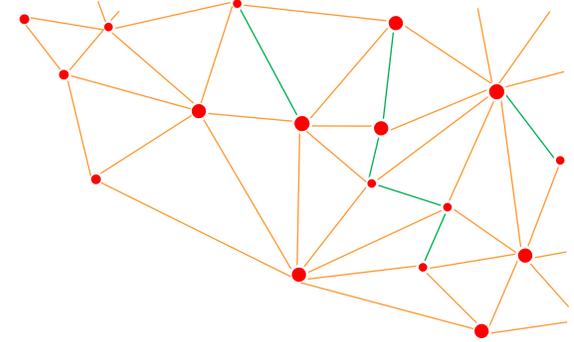
2. EWN



3. ВTB-WN

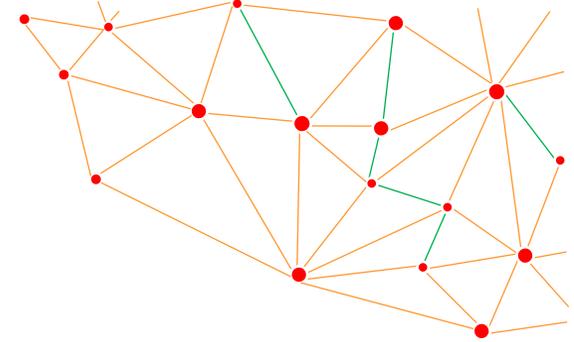


Missing Concepts in EWN



The gaps that we observe in EWN are of two main kinds:

- Language differences between English and Bulgarian
- Due to differing approaches

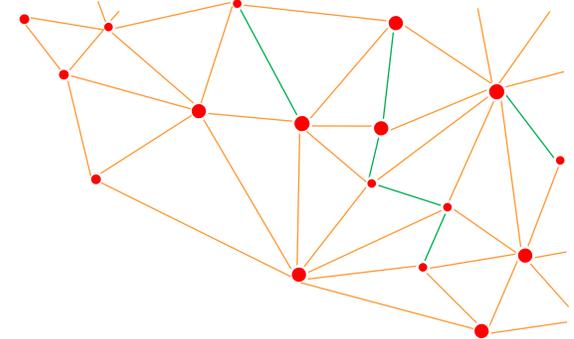


Missing Concepts in EWN

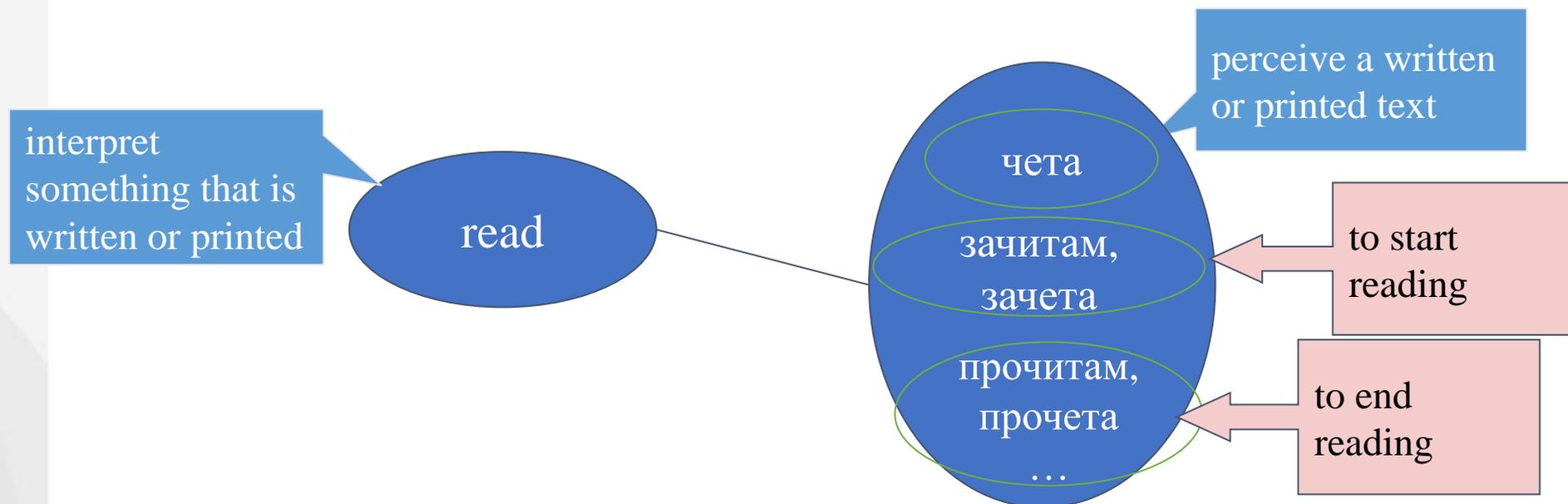
- Natural differences between English and Bulgarian:
 - concept differences
 - Bulgarian reflexive verbs

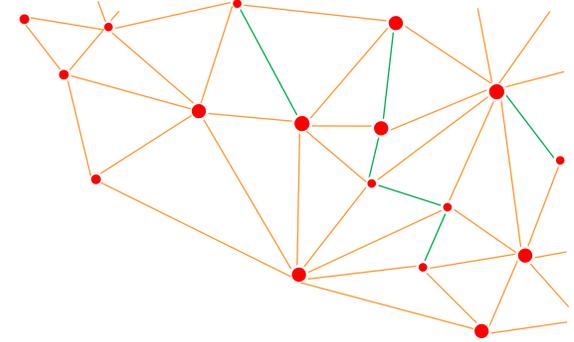


Additional Linguistic Information



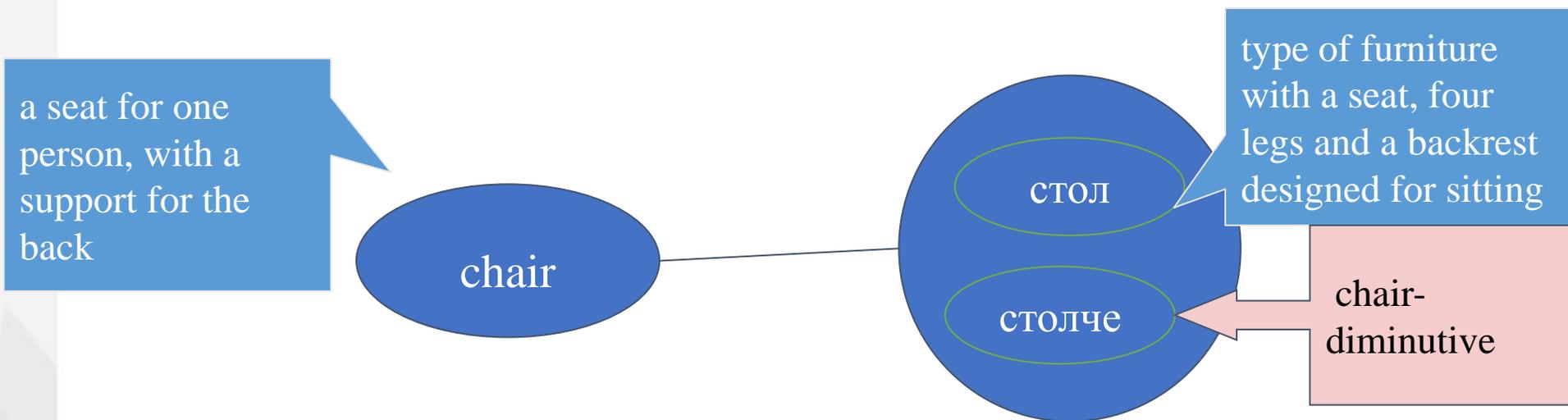
Bulgarian verbs with prefixes that bear semantics of start, end, duration, repeatability, etc. of the action can not have an equal English synset so they will be mapped with the general meaning of the verb and labeled with their specific semantic features on the level of lemma

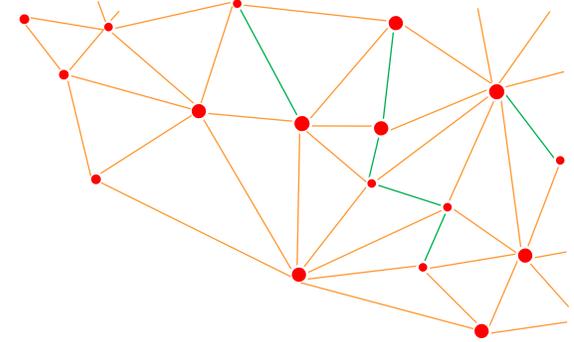




Additional Linguistic Information

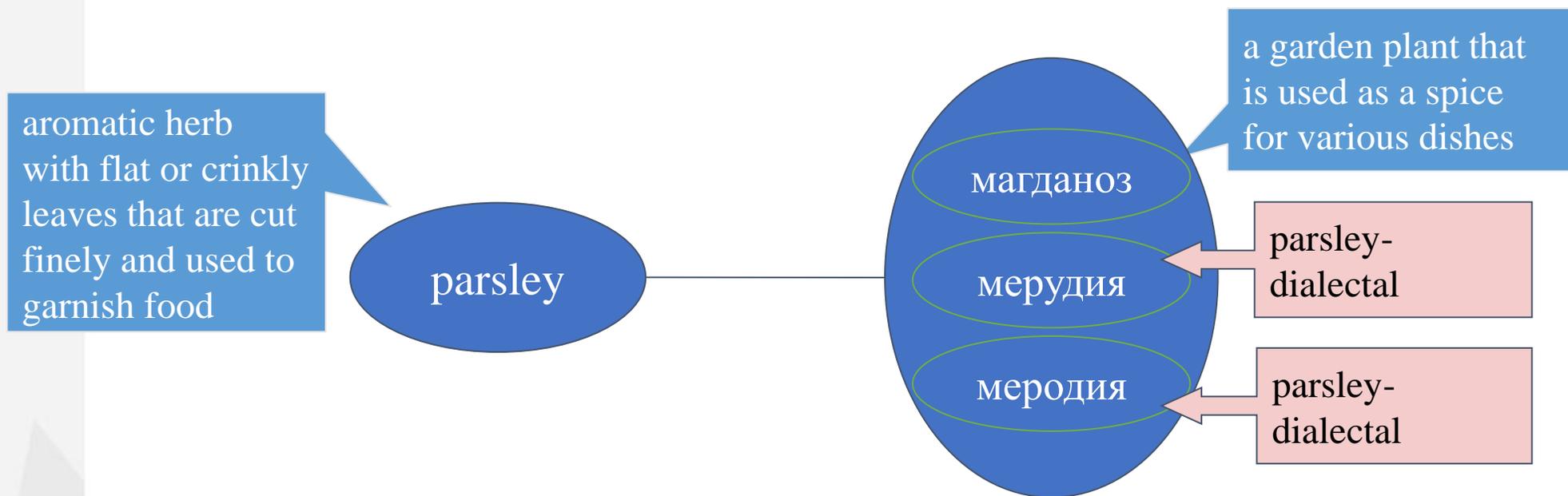
The diminutive forms of the nouns will be in one synset with the general form. Bulgarian diminutives can have more than one meaning. The general one is that something is very small or very young, but they can also express gentle, diminishing or humiliating attitude.



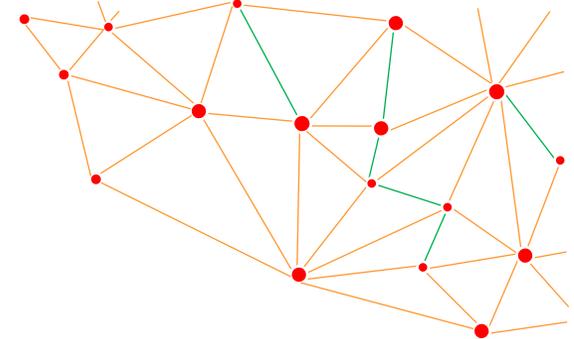


Additional Linguistic Information

Lemma markers for archaic, dialectal, slang, informal, vulgar, offensive, etc.



Additional Linguistic Information



Marker for MWE lemmas

an anniversary of the day on which a person was born (or the celebration of it)

birthday

рожден ден

anniversary of the date a person is born

multi-word expression

without the use of a machine

by hand

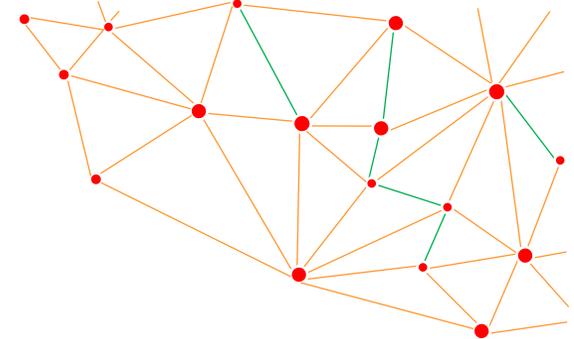
ръчно

на ръка

by hand rather than using a machine or other mechanism

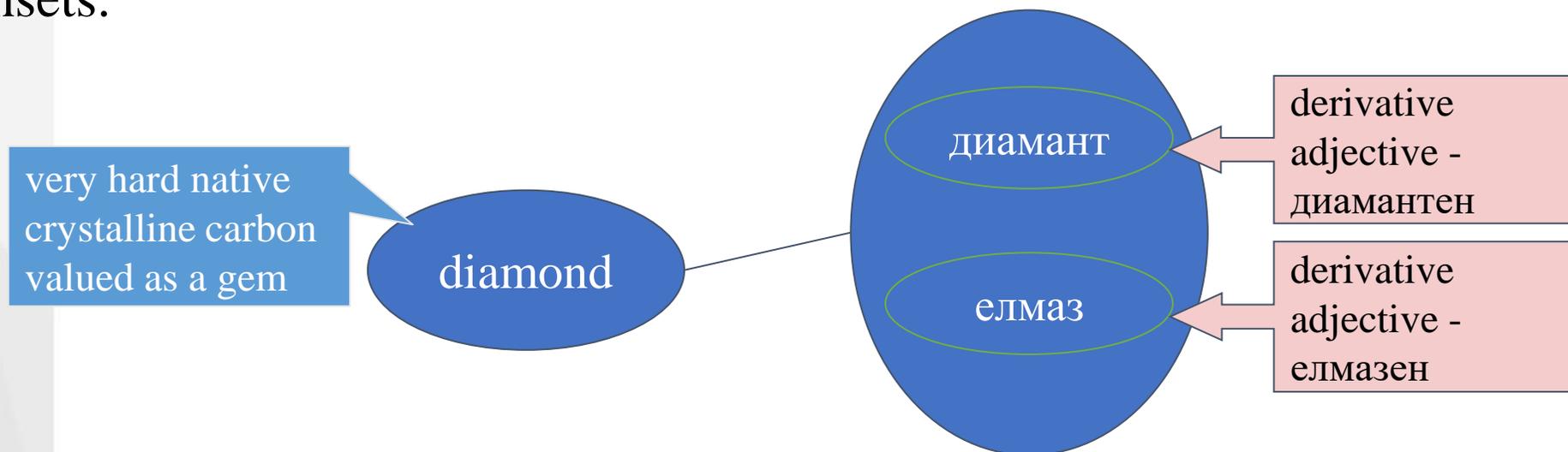
multi-word expression

Additional Linguistic Information

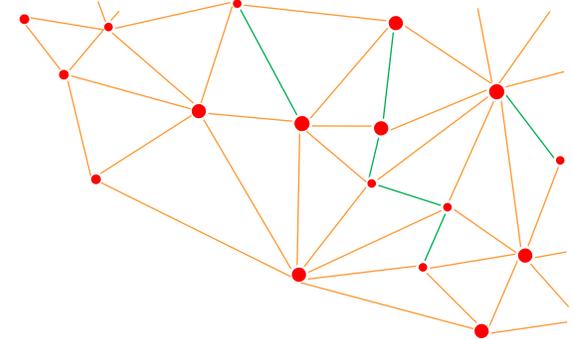


Derivational relations between different parts of speech

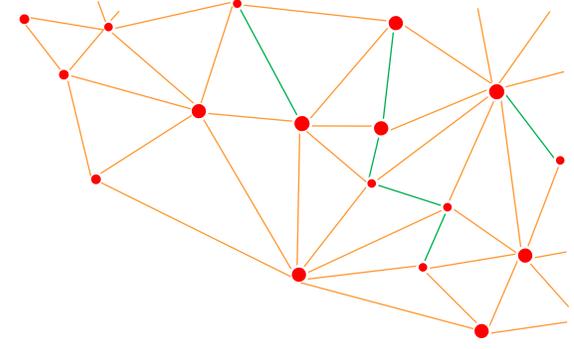
These markers are again used for lemmas rather than synsets because if they are applied on synsets they would not be appropriate for members of one synset that are not derivationally related. A very common example is the conversion of English nouns to adjectives. Thus, for many noun synsets there are no adjective synsets.



Mid-Observations

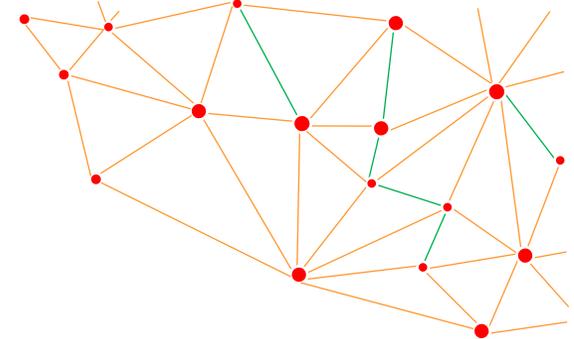


- The challenges that are encountered in the mapping of two wordnets are related with the natural differences between the two languages but also depend on the way that the resource is built.
- The discrepancies are overpowered with interlingual relations and relations between lemmas and synsets in the BulTreeBank WordNet.



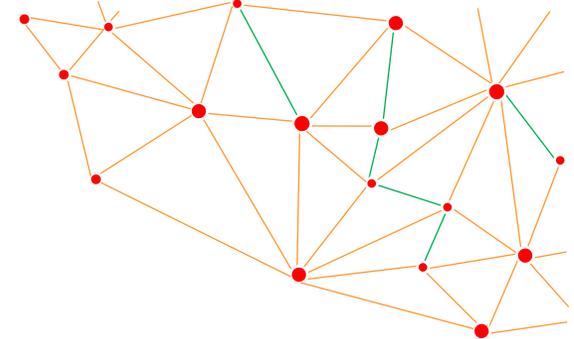
BTB-WordNet as a Hub for Resource Integration

Introduction



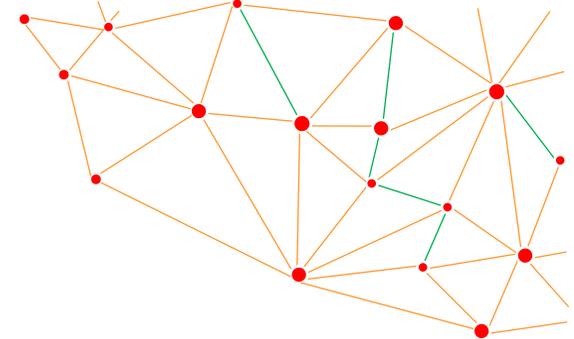
- The reported work here refers to the last three years (2020, 2021, 2022)
- It turned out that many NLP applications required **not only available resources but also appropriate integration among them**
- We started to view **BTB-WN** as a hub for linking grammar, other lexical data and world knowledge
- 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



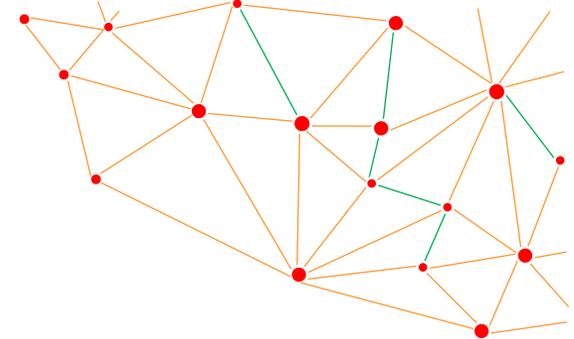
- The development of **BTB-WN** goes back to the times when an *Ontology-based lexicon for Bulgarian* was initially constructed (**Simov and Osenova 2010**)
- Here we started with domain ontologies aligned to the upper ontology *DOLCE*, using *OntoWordNet* for introducing the middle level concepts
- 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 (<https://omwn.org/>)

Some Broader Context



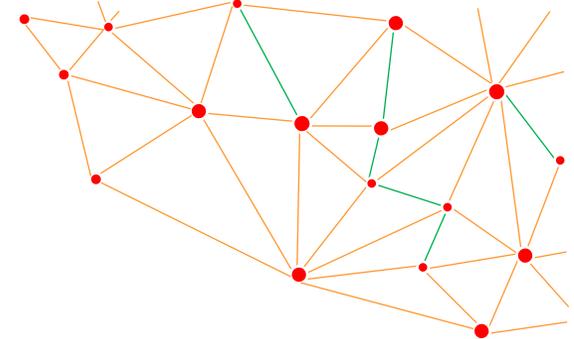
- The interface between lexical semantics and grammar, between lexicons and corpora has been extensively discussed from various points of view: linguistic, typological, formal, implementational, etc.
- We support the point of view in which the grammar is born in the lexicon, i.e. the *lexicalist-centric 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



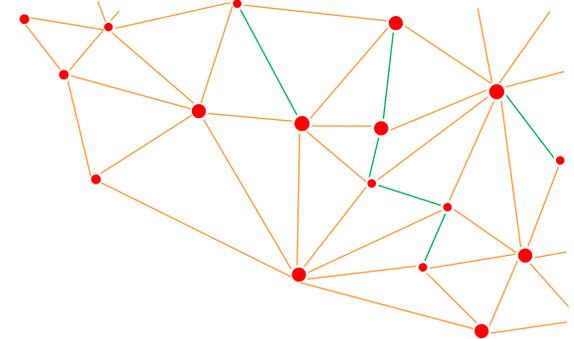
- It is well-known that **wordnets are thesauri**. Despite providing the meanings of words grouped within synsets and relations among these synsets, they are still:
 - very static
 - self-contained and
 - often do not cover all parts of speech
- At the same time, they are good candidates for playing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

Linking of BTB-WN



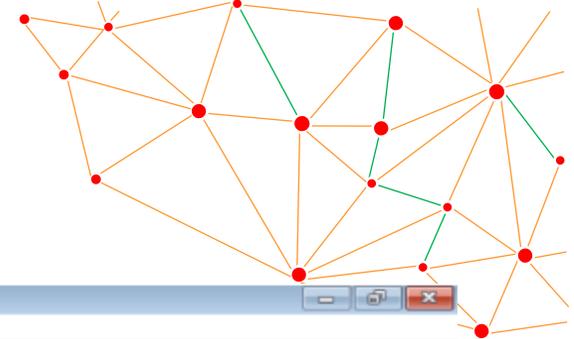
- We switched from a tool that supported only local editing (where synsets were considered within a very limited context) ([CLaRK system](#)) to a tool that supports editing of the Wordnet data within a global context ([CLaDA-BG-Dict](#))
- When a lemma is selected within [BTB-WN](#), the following information 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



CLaDA-BG-Dict

Действия Форми Инструменти Помощ История

Списък от лемите

Начало на лемите: Част на речта: Намерени са 120 лема. (0.354*)

Категория: Релация:

Само лемите с отворени въпроси

Темата на въпроса да съдържа низа:

Въпросът да съдържа низа:

№	Лема	Част на речта	Еквивалент	Тикет
97	късо съединение	п	Е	
98	късоврат	а		
99	късовълнов	а		
100	късоглед	а	Е	
101	късоглед	п	Е	
102	късоглед	с	Е	
103	късогледство	п	е	
104	късокрак	а		
105	късопаметен	а		
106	късопръст	а		
107	късче	п	Е	
108	кът	п	Е	
109	къче	п	Е	
110	къч	п	Е	
111	къшей	п		
112	къшпа	п	Е	
113	къшлак	п	Е	
114	къща	п	Е	1
115	къщен	а	Е	
116	къщи	г	Е	
117	къщица	п	Е	
118	къщичка	п	Е	
119	къщурка	п	Е	
120	къщя	п	Е	

Лема: къща

Синонимно гнездо

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

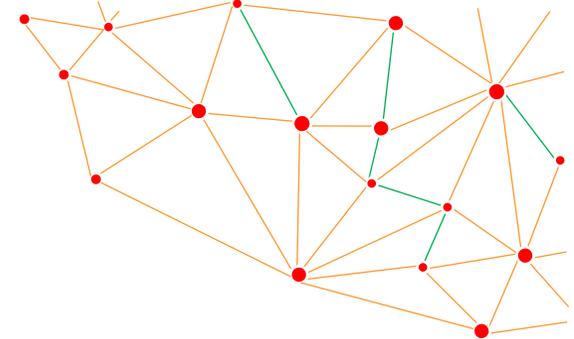
Лексикална единица

3 примера, 3 от които към лема

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

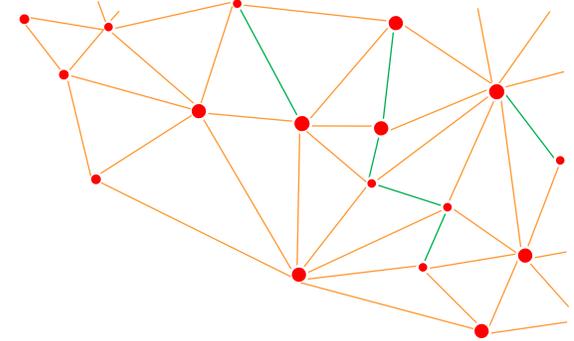
Концептуални релации

Linking of BTB-WN (3)



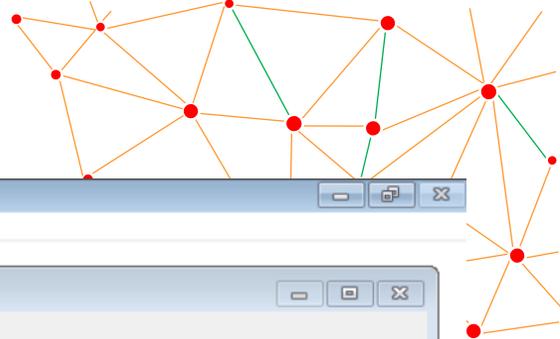
- In addition to granting access to OEW, the system provides access to dictionaries that are freely available to us, among which the **Bulgarian Explanatory Dictionary**, our **in-house Bulgarian Inflectional dictionary**, two **Bulgarian-English dictionaries**
- Each of these dictionaries could be consulted 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 лема

Начало на лемите: ма

Категория:

без

Само лемите с от

Темата на въпроса д

Въпросът д

Лема: вълшебник

Синонимно гнездо

Част	Категория
n	noun.person

Лексикална единица

5 примера

Релации Надпоятия

Човек, който притежава въздейства на природата

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

Търсене на дума в дефиниците/примерите

Израз за търсене: вълшебни

Търси

Нов пример

Различаване на големи и малки букви Търсене като низ

Добавяй резултатите от търсенето към предишните

В дефиниците В примерите

В корпус BGLITPLUSCORPUSWN.DBF_CORPUS

Изходен текст	Елемент	Десен контекст
добри	вълшебни	ци.***BR***
рът за	вълшебни	ци.***BR***Дълго се вглеждаш в дълбините на т
стана,	вълшебни	ци? Напълнихте ли гушките?!***BR***Гудвини си
в няма	вълшебни	ци?***BR***— А защо Свирулкин казва, че е вид
амо за	вълшебни	ци?***BR***— Първо на първо, да те светна, че
ни като	вълшебни	ци...***BR***Бяха дошли и в техния град. Не на і
вар за	вълшебни	ци". Каниш се да я разгърнеш, но господин Едог
вар за	вълшебни	ци". Каниш се да я разгърнеш, но господин Едог
вар за	вълшебни	ци". Тъкмо като за тебе, решаваш ти и се заглеж
ата на	вълшебни	ците Глъбдъбдриб. Тук пред него оживява в чое
*BR***	Вълшебни	ците и вълшебствата***BR***Гледах видеото им
т само	вълшебни	ците и никой да не смее да им се подиграва! А,
лно от	вълшебни	ците на краля — странно, но уютни шатри от спл
ена от	вълшебни	ците на перото.***BR***Келнерът Ибрахим се п
о бяха	вълшебни	ците на светлите елфи. Те спуснаха сияен щит, і
*Само	вълшебни	ците не остаряват***BR***Когато 22-годишният
. Само	вълшебни	ците не остаряват***BR***Само вълшебниците
гаш, че	вълшебни	ците са глупост?***BR***Незайко взе да разка
куси, а	вълшебни	ците са измислица. Каквото си направил сам, н
война	вълшебни	ците се биеха помежду си заедно с воюващите :
ругаде	вълшебни	ците си, защото ти си роден в пъкля и демон о
и, ние,	вълшебни	ците сме като всички други. С годините придоби
и те, и	вълшебни	ците ще съществуват, докато някой им върва и с
аваш с	вълшебни	ците! — строго каза Карфичка. — Никой не мож
авки са	вълшебни	ците!***BR***— Разбира се, че не зная — сви р:
ду нас,	вълшебни	ците, не бива да има прегради...***BR***Кабине
ците,	вълшебни	ците си казва те и в миг изпита толкова сила пр

В документ: C:\WORK\PROJECTS\BG-CLARIN\2022\MATERIALS\WORDNET2020\WORDNETEDIT...
 ои късмет тъкмо тогава, забелязах в тъпата 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 наред Дъбовридския пазар, без жезъл! Направо ще предизвикате повече смях и радост сред пъстрата тъпа от таласъма Нийл Смешника!
 — Здравсти! — 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 реч	Категория	Дефиниция
n	noun.person	Човек, който притежава с

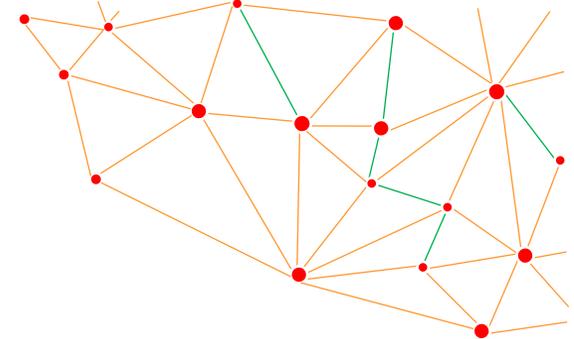
Лема

вълшебник

Прикачване към лемата

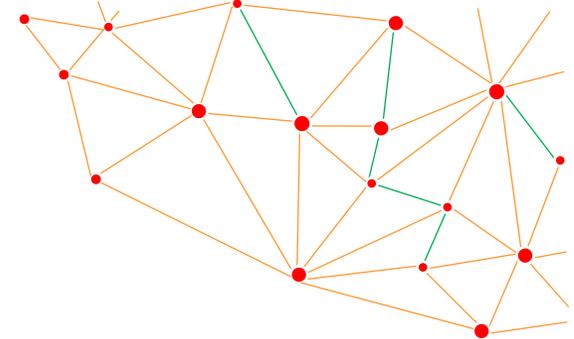
Добави примера

Link to Valency Dictionary



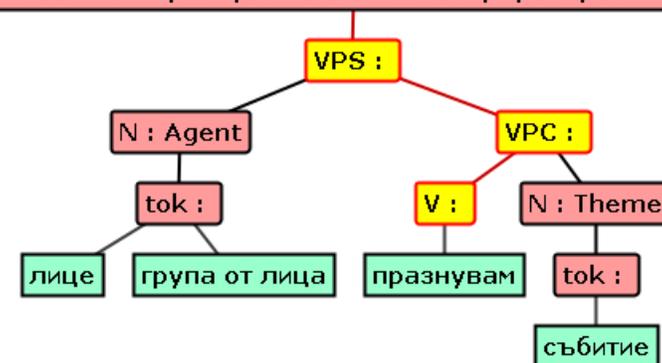
- The verb in a valency frame is connected to **BTB-WN** via a mapping to an appropriate synset from where access to the lexicographic class (such as *verb.social*, *verb.cognition*, etc.), the list of lemmas and the definition are available
- For example, if the *verb.emotion* **worry** is considered, 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

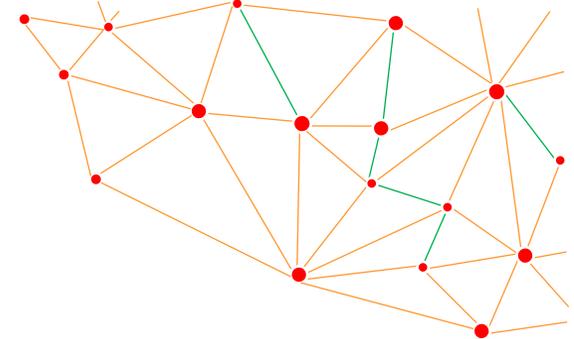


- ▢ FramesDef: :лице група от лица празнувам събитие
 - ▢ FD: VerbNet:judgement-33
 - ▢ lemma: празнувам
 - ▢ def : Чествам, прекарвам някой празник.
 - ▢ F: verb.social LEMMA: празнувам DEF: Чествам, прекарвам някой празник.
 - ▢ VPS: :> лице група от лица празнувам събитие
 - ▢ N: Agent :> лице група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

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

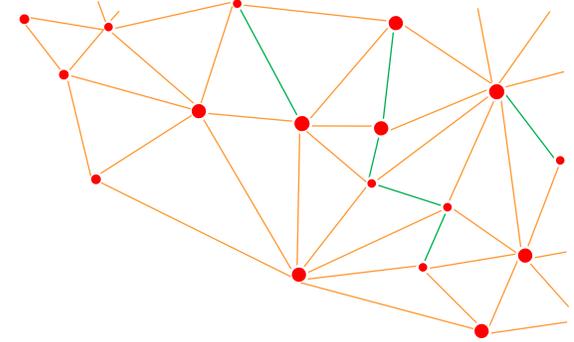


From CLaRK to CLaDA-BG-Dict



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:

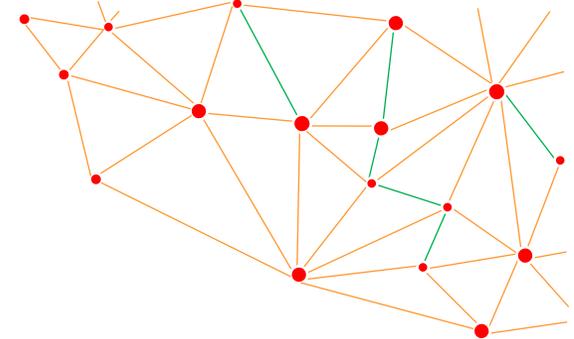
- Appropriateness of definitions
- Alignment to OEW
- Missing senses
- Wrong or missing relations
- Appropriateness of the examples



Appropriateness of Definitions

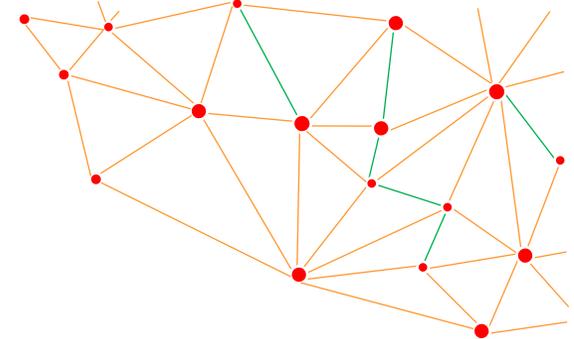
- We checked the definitions for the different kinds of word classes per synset
- This step was necessary, because we wanted to extend the definition to include more information than the definition within paper dictionaries
- This holds especially for adjectives. In the traditional dictionaries the adjective 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

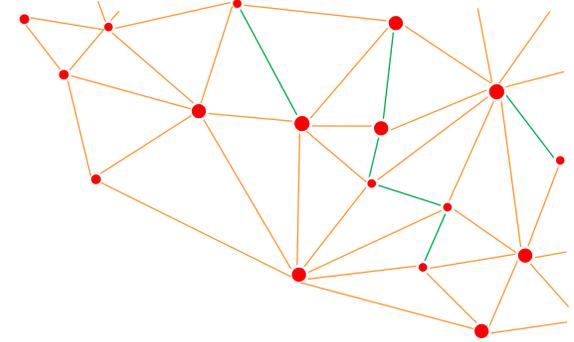


- In the previous versions having only a local view, we supported as many relations as possible between the Bulgarian and the English synsets some of which allowed in the noun and verb hierarchies to have disconnected elements
- With the switch to the global view it became much more convenient to verify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

Missing Senses

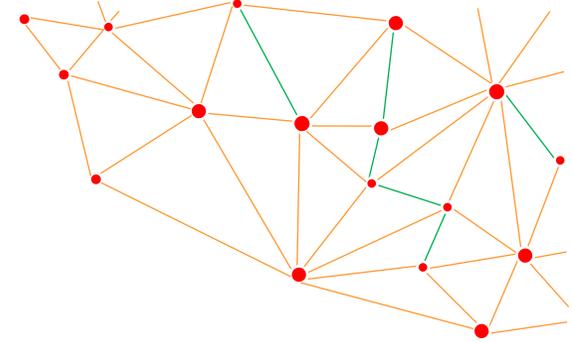


- The construction of the early versions of **BTB-WN** were mainly driven by specific NLP tasks like Word Sense Disambiguation, Machine Translation, Mapping to Domain Ontologies
- In this applications we had to cover certain domains or type of texts. This resulted in representation of the senses of the 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

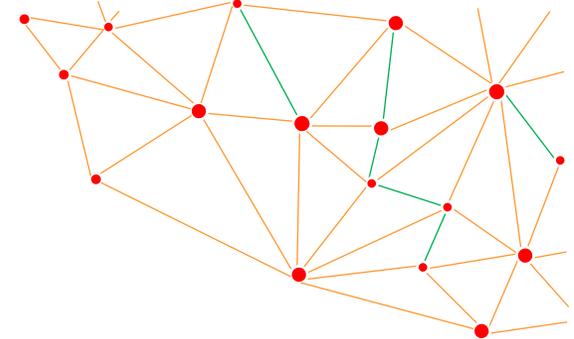
- From the beginning we supported mapping to PWN (and now EOW). We are using this mapping to transfer automatically relations from English synsets to Bulgarian ones
- After the transfer the set of relations became eligible for modifications, if needed. This happens mainly when the mapping is not between equivalent synsets



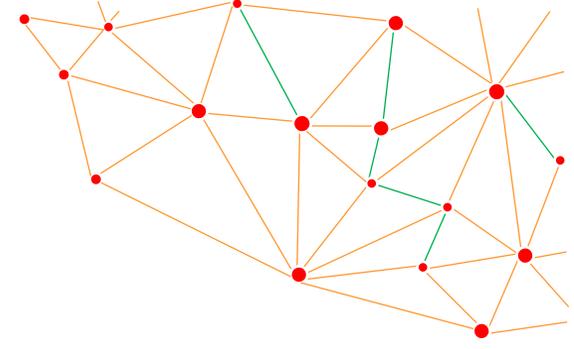
Appropriateness of Examples

- 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

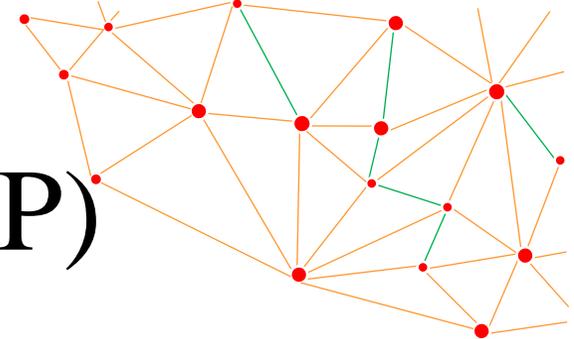


- Besides the examination of the existing synsets we extended **BTB-WN** with new synsets through the above mentioned vocabularies extracted from both types of sources - dictionaries and corpora
- Then the following information was added: derivational sets for these lemmas such as adjectives derived from nouns, aspectual variants of Bulgarian verbs that share a common basic sense, etc.
- In this way, more than **16 000 synsets** were added (in total **35 000**)
- At the moment we completed the coverage of the core vocabulary with about **6000** lemmas.



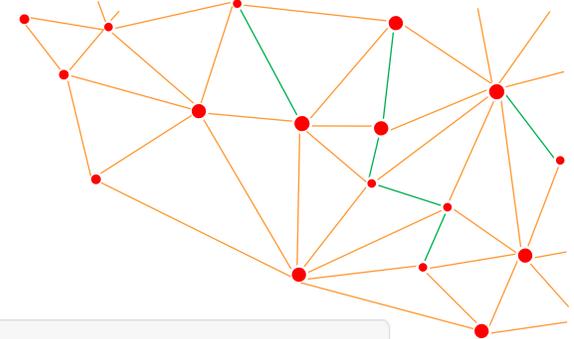
BTB-WordNet based applications

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

ВТВ-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. ↑

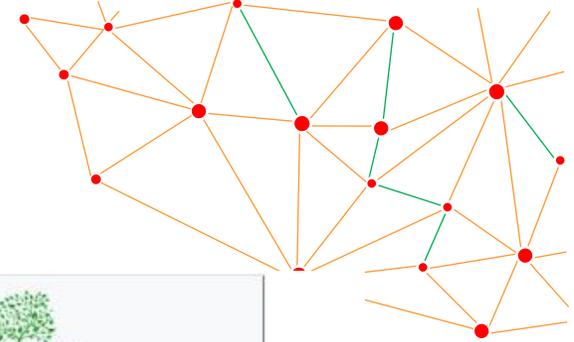
Документ за получени или изплатени суми (срещу продадена стока или извършена работа, услуга). Документ, сметка с подробно описание на продадена/купена стока. ↑

Лична изгода, облага от нещо. ↑

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

Missing part from the user interface: relations to other dictionaries

ВТВ-Wordnet Viewer



CLARKE Българска мрежа от думи - ВТВ-WordNet BulTreeBank

ябълка

търси форми

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

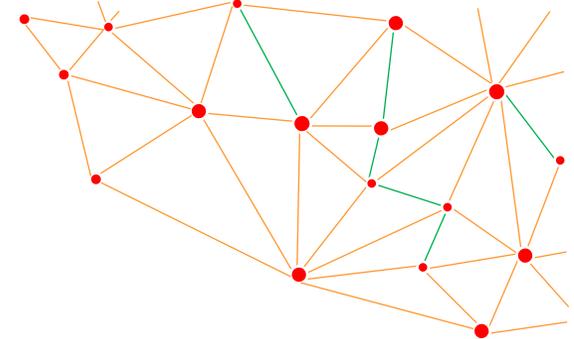
→ (п) ябълка
Овощно дърво с розови цветове и възкисел или сладък сочен плод с кълбовидна форма. ≈ **apple tree**
*Дъждът дойде, но подранила есен от **ябълката** късаше листа.*

→ (п) 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.*

```
graph TD; ябълка -- hyr --> дръвче; ябълка -- hyr --> овощка; ябълка -- sdt --> ябълков; ябълка -- eqt --> apple_tree[apple tree]; ябълка -- hys --> дива_ябълка[дива ябълка];
```

Game of Meanings

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



CLaRK

BG EN

Игра на значения (beta)

BulreeBank

9 / 10

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

Упражнения по българска граматика

Множествено число #1

1 2 3 4

- Извинявай, дай една игла.
- Ти имаш много ? .

игла
иглата
игли
иглите

Модел:
- Извинявай, дай един молив.
- Ти имаш много моливи.

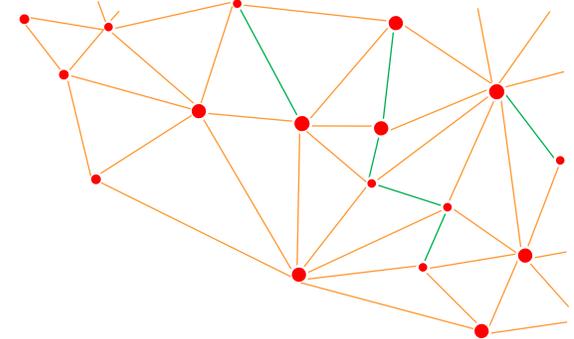
УПРАЖНЕНИЯ: (beta)

- 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)

CLaDA BG CLARIN DARIAH-EU МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА ИИКТ

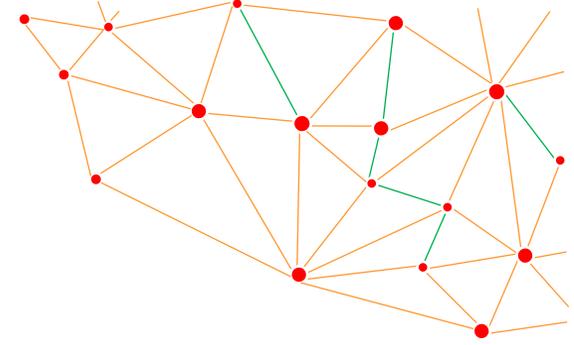
Using BTB-WN and Valency Dictionary to impose semantic constraints on suggested words

Background: Motivation



- Lack of sufficient knowledge for solving many important NLP tasks, such as WSD, Relation Extraction, Entity Linking, Semantic Annotation, etc.
- Number of integrating efforts:
 - SemLink
 - PredicateMatrix
 - Uby
 - BabelNet

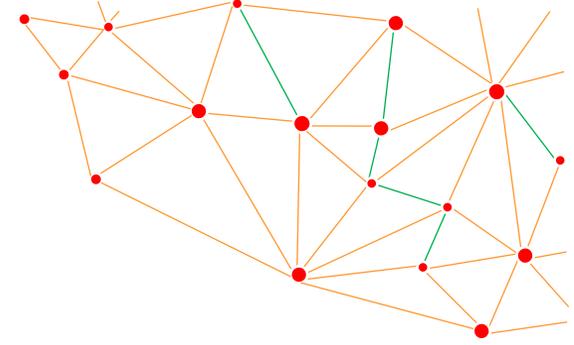
Background: Motivation



- Two facts are demonstrated:
 - A single knowledge resource is not sufficient for the most of the NLP tasks
 - The automatic integration of the various distinct resources is error prone

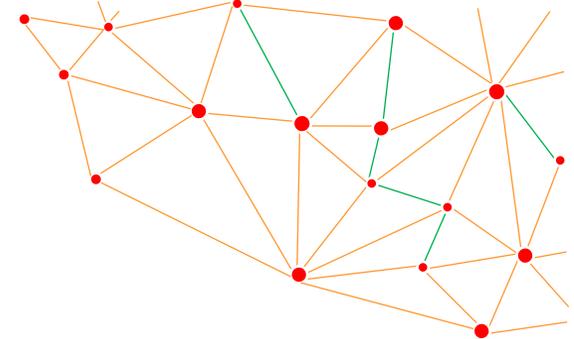
(This is especially true for low-resourced languages in semantics.)

Background: Tasks



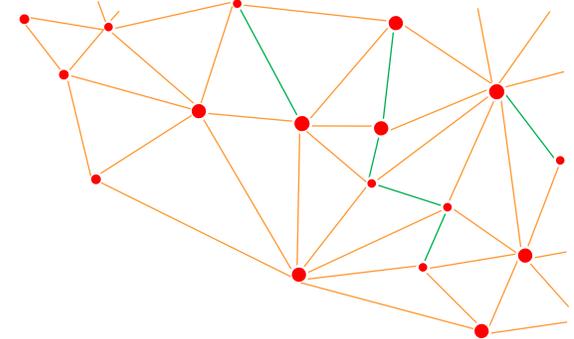
- Focus on the integration of BTB-WN and Wikipedia
 - (in general) Mapping of concepts in WordNet (synsets) to concepts and instances (named entities) in Wikipedia
 - The result is a new version of BTB-WN extended with:
 - New senses and new synonyms for the existing synsets
 - A controlled number of named entities that are specific to Bulgaria
 - Increasing the number of terminological concepts

Background: Tasks



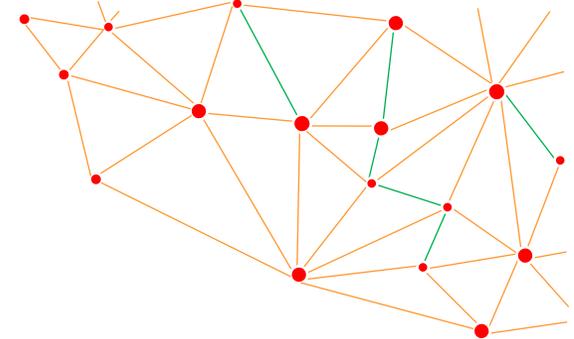
- The work has been done manually, BUT with an automated preparatory phase
- The integrated resource combines general lexica with encyclopedic knowledge (terminology)
- The expected result would be twofold:
 - Mutual enrichment and improvement of both resources
 - An integrated resource that provides access to knowledge graphs (DBpedia, Wikidata – via Wikipedia URIs)

Related Work: BabelNet



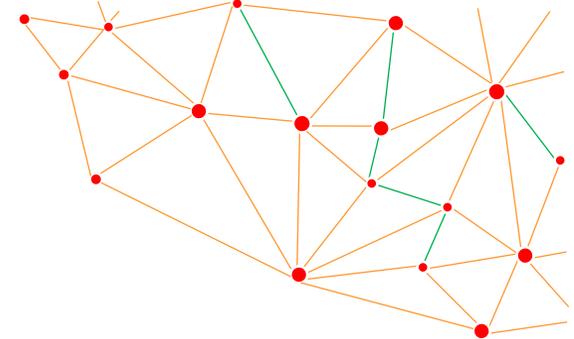
- BabelNet - an automatically created very large, wide-coverage multilingual semantic network
 - Encodes knowledge as a labeled directed graph
 - Created by linking the largest multilingual Web encyclopedia – Wikipedia, to the most popular computational lexicon – WordNet

Related Work: BabelNet



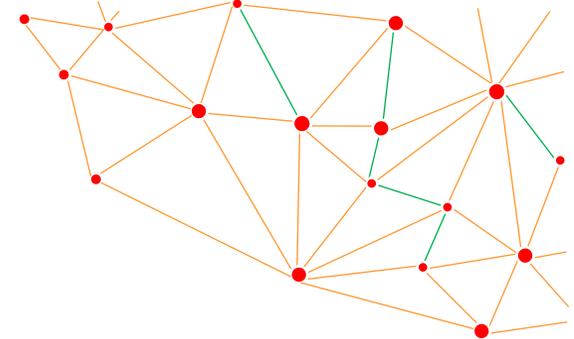
- Relation to our work:
 - Adding more **locally important** content into the existing mappings
 - Enriching the resource that was constructed automatically with **validated data**
- The *Babelfy service* is very good at detecting concepts and names, but it still has problems with disambiguation among local (Bulgarian) people or places with the same name, or between a concept and a name

Related Work: BabelNet



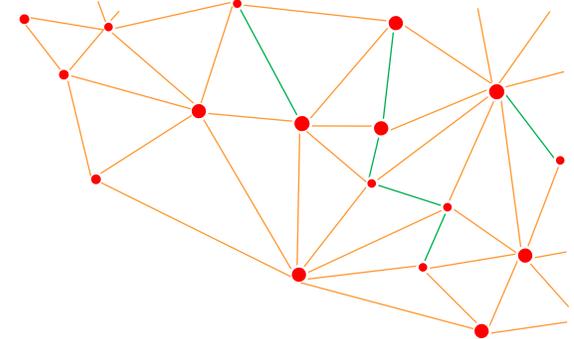
- For example, the verb **литва** (*litva, start to fly*) is identified only as the country **Литва** (*Litva, Lithuania*) whose graphical form coincides with the verb
- Similar for the adjective **русия** (*rusiya, blond*) and the name of the country **Русия** (*Rusiyā, Russia*)

Related Work

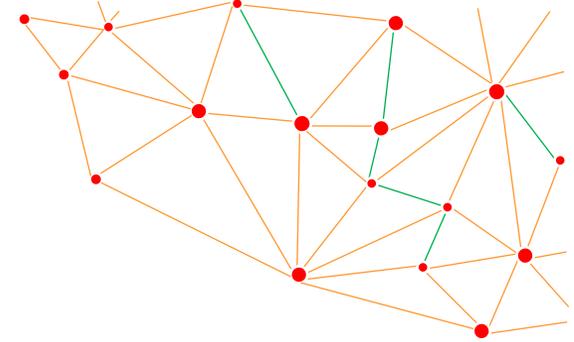


- (Osenova and Simov 2018) mention the initial attempt for annotating of named entities (NE) in BulTreeBank with URIs from DBpedia, to have access to instance information:
 - **However:** the BulTreeBank appeared to contain only a small number of named entities in Wikipedia
 - Thus the extension was insufficient and it required the use of the Wikipedia URIs and DBpedia classes for the missing NEs

Related Work

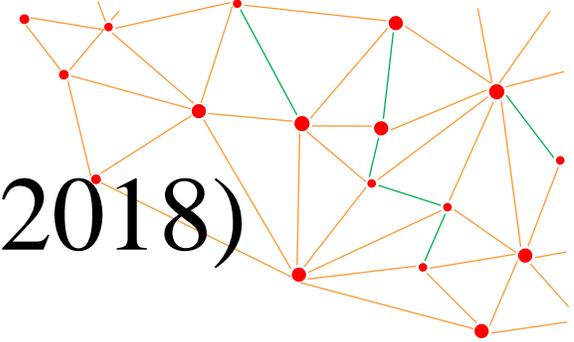


- **(Rudnicka et al., 2017)** present another attempt at linking two large lexico-semantic databases - Princeton WordNet and the plWordnet. The approach considers models and ideas originating from the bilingual lexicography and translation studies
- A recent and rather innovative example of the development of a FrameNet based on a corpus of written Dutch, and annotated with PropBank predicates and roles is the project of **(Vossen et al., 2018)**



Related Work

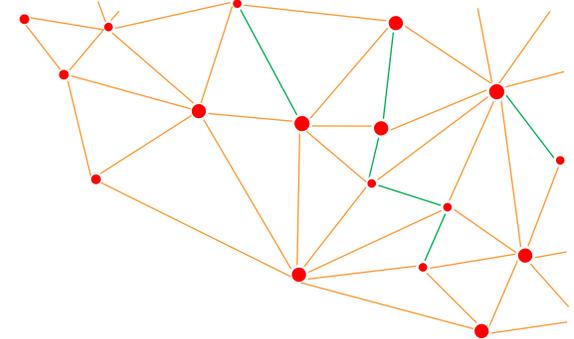
- (McCrae, 2018) reports on the mapping of the **Princeton WordNet (PWN)** instances to the **English Wikipedia**
- A subset of PWN instance concept synsets is automatically linked and manually evaluated on Wikipedia articles in order to *provide a gold standard for link discovery*



Typology of Mapping Links (McCrae, 2018)

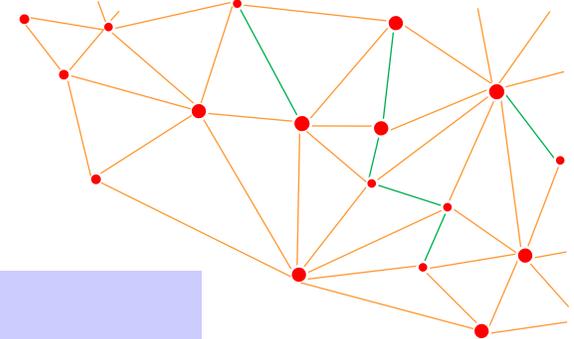
- 5 types of links between PWN and Wikipedia:
 - **Exact:** one synset to one article
 - **Broad:** several synsets to one article
 - **Narrow:** one synset to several articles
 - **Related:** one-to-one relation, but not the same concept
 - **Unmapped:** not possible to map
- We adopt this typology but we aim to map concepts in addition to instances

Mapping ВТВ-WN to BG Wikipedia



- Preparatory automated work:
 - For each lemma within ВТВ-WN all matching articles in Wikipedia were automatically selected
 - The article titles were cleaned in advance from the modifiers given in brackets
 - The lemma **маса** as:
 - a) **Маса_(мебел)** masa (table, a piece of furniture),
 - b) **Маса (величина)** mass (a body of matter), and
 - c) Maca mob (a disorderly crowd of people)
 - Is mapped to Wikipedia articles with the appropriate titles

User Interface



```
eq : Айкидо
eq : Айнджовен
eq : Айнщайн
eq : Айнщайний
eq : Айова
eq : Айсберг
  page : Айсберг Iceberg : ''Айсберг'' ({{lang|de|Eisberg}}, буквално означаващо „ле
  entry : 09308572-n :Айсберг=: : : : айсберг > Огромен леден блок, откъснал се
    title : Айсберг :
    cwn : {09331478} <noun.object>[17] S: (n) iceberg#1 (iceberg%1:17:00::), berg#1 (ber
    bg: айсберг
    senses : Огромен леден блок, откъснал се от полярен ледник, който плава или лежи на
eq : Академия
eq : Акари
eq : Акация
eq : Акварел
eq : Аквариум
  page : Аквариум :Aquarium:: ''Аквариумът'' е съд, предназначен за отглеждане на [[p
  page : Аквариум (група) :Aquarium (band):: ''„Аквариум“'' от [[Санкт Петербург]] е
  page : Аквариум (пояснение) :*** disambiguation page ***: ''Аквариум'' може да се
  page : Аквариум (филм, 1895) :: ''"Аквариум"'' ({{lang|fr|Aquarium}}) е [[Франция|
  page : Аквариум (филм, 2009) :Fish Tank (film):: ''„Аквариум“'' ({{lang|en|Fish Tan
  entry : 02732072-n :Аквариум=: : : : аквариум > Съд, обикновено стъклен, пълен
eq : Акведукт
eq : Акче
```

Bulgarian Title

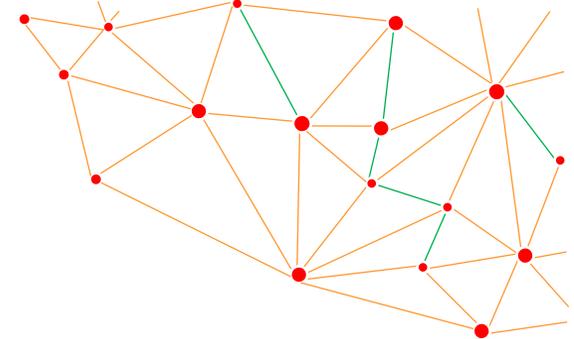
English Title

The beginning of the article

One or more Wikipedia pages

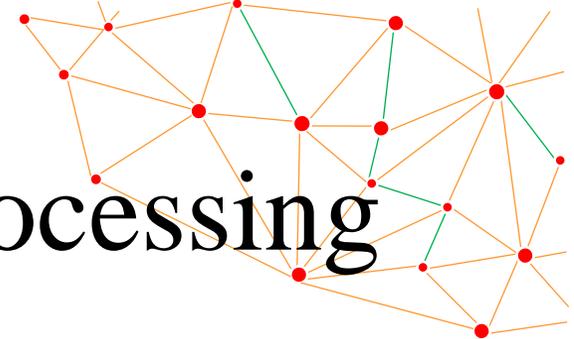
One or more synset

Mapping



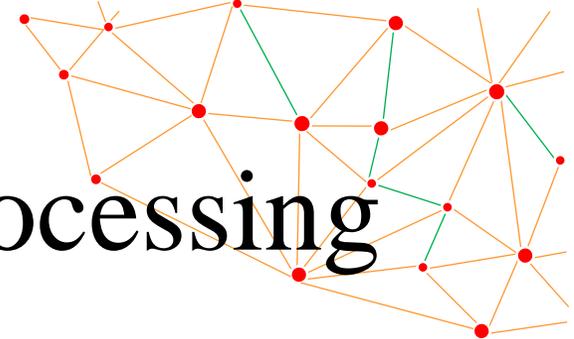
- For about **22 511 synsets** in BTB-WN, **27 984 lemmas**, **44 873 senses**
- A little more than **13 000 Wikipedia articles** have been extracted
- Apart from the mapping procedure, after consulting the individual senses in BTB-WN, the annotators checked whether new meanings had to be added to it
- The new meaning could be **a sense** for the **common word** or a **named entity**
- In both cases the annotator created a new synset entry in BTB-WN

Instance Mapping: Named Entities Processing



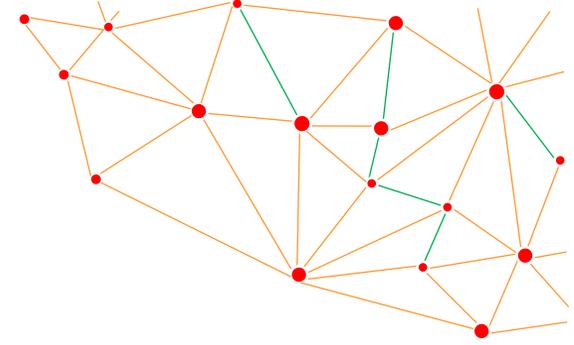
- Our approach:
 - Due to the high productivity in the case of named entities, many common words are presented as named entities in Wikipedia
 - We aim at Bulgaria-centered mappings
 - The annotator first filtered the candidates in order to introduce only the important names:
 - As a first step, only names of persons, organizations and locations are considered
 - For location names we select names of Bulgarian places or of well-known foreign places
 - For the rest of the names only well-known names are considered

Instance Mapping: Named Entities Processing



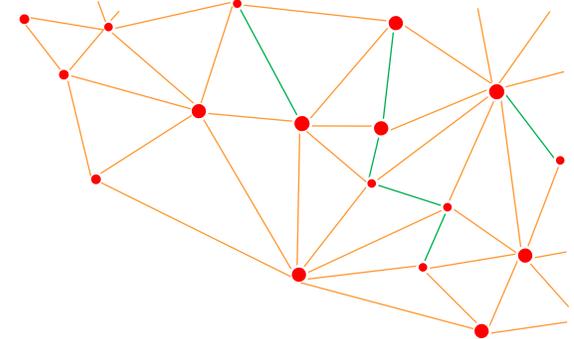
- Initially, a restriction to include larger cities in the world was introduced (**larger than 100 000 citizens if they are not well-known**)
 - **Шенген** (Schengen) is included in BTB-WN although it has less than 4000 citizens, but
 - **Буден** (Boden, a city in Sweden) is not included although its transliteration in Bulgarian coincides with an adjective (=awake)

Discussion



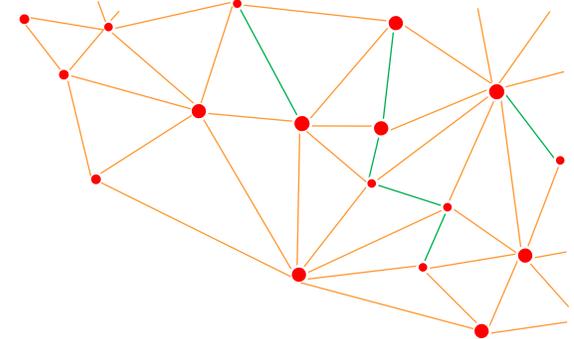
- It is evident that the above selection criteria are more or less **arbitrary**
- In our future work we need to make the definition more precise in order to cover all the names in Wikipedia, but
 - Without overloading the WordNet with the ambiguity coming from very rare named entities.
- At the moment we use a gazetteer as an initial filter

The Utility of the Gazetteer



- All the Wikipedia pages that correspond to the names in the gazetteer were extracted (**10 899 pages**). From them **1 515** already in BTB-WN
- The rest **9 384** pages were classified as:
 - Bulgarian locations
 - Other locations
 - People
 - Organizations and
 - Other
- They will be checked for inclusion in BTB-WN at a later stage
- In this way we selected also some important names that are not considered at the beginning of this work

Cases of Mapping

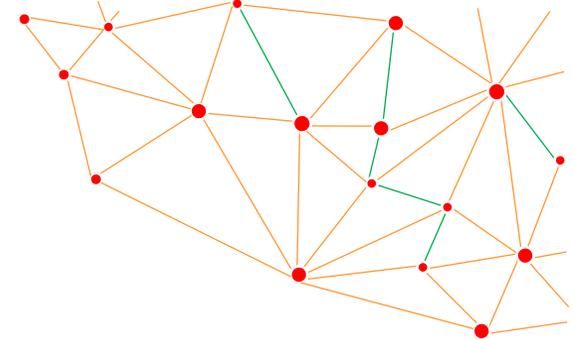


1. **Exact mapping** of senses represented in both resources
2. A **concept represented in Wikipedia, but not in WordNet**
3. A **named entity in Wikipedia, missing in WordNet**

In cases 2 and 3 annotators had to create a new synset and to establish a mapping

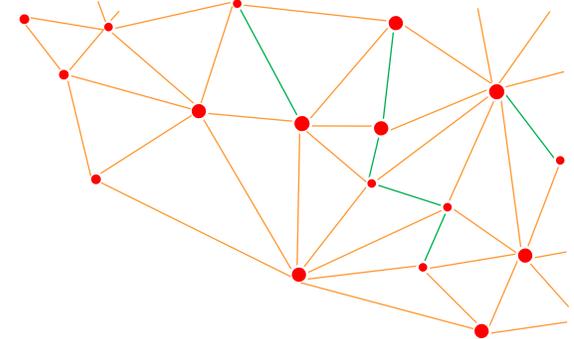
The annotation was performed by **5 people** who considered nearly **1000** WordNet lemmas, automatically mapped to more than **13000** Wikipedia articles

Statistics over Current Mappings



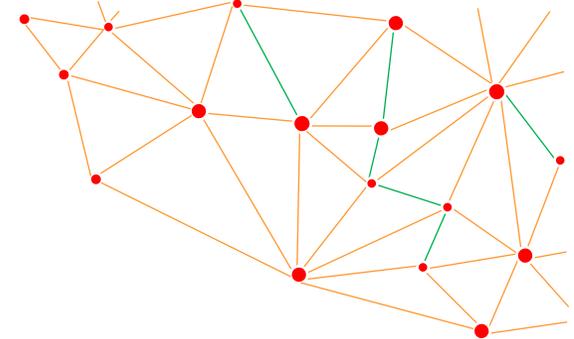
Correspondence	Number	%
	Total: 1309	
None	276	21.08
Equality	688	52.57
Many to One	128	09.78
New Concept	128	09.78
New Named Entity	68	05.19
New Synonyms	21	01.60

Examples: No mapping



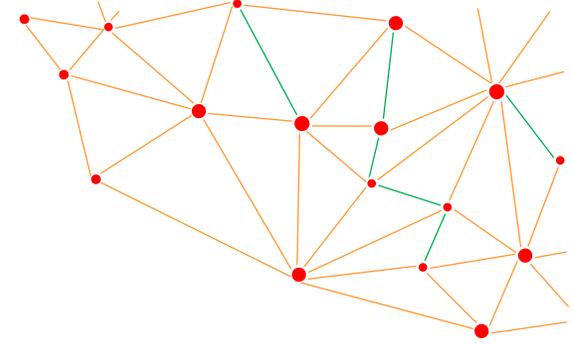
- Стожер (stozher) in the Wikipedia is only a name of *a village* and *a newspaper*, while WordNet records only the concept **стожер** (stozher) as *pillar* (missing in Wikipedia)
- Thus, the WordNet entity cannot be mapped to Wikipedia. This case corresponds to McCrae's *Unmapped links*

Examples: Equality



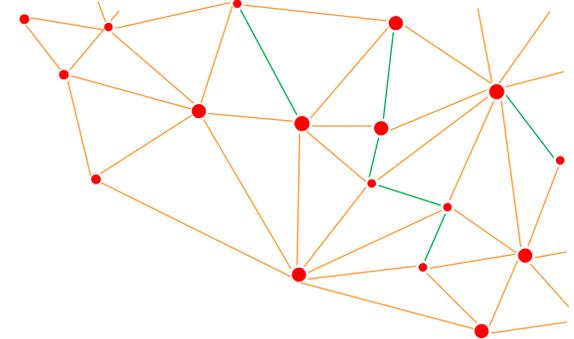
- **Столица** (stolitsa, capital) is defined in the same way in both resources
- These cases are the majority of all mappings. It corresponds to McCrae's *Exact links*

Examples: Many-to-One



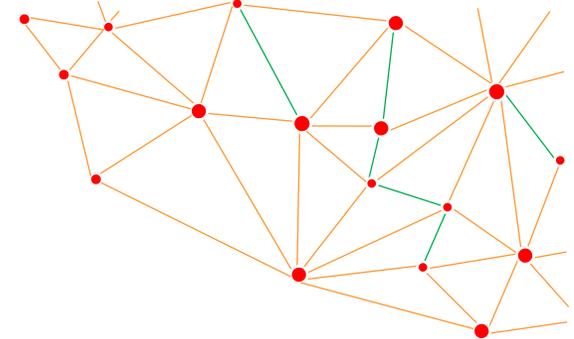
- Different parts of the same Wikipedia article are dedicated to different concepts
- Often this is the case for the disambiguation pages. Among the concepts, one usually corresponds to the mapped WordNet synset
- Corresponds to McCrae's *Broad links*

Examples: Many-to-One



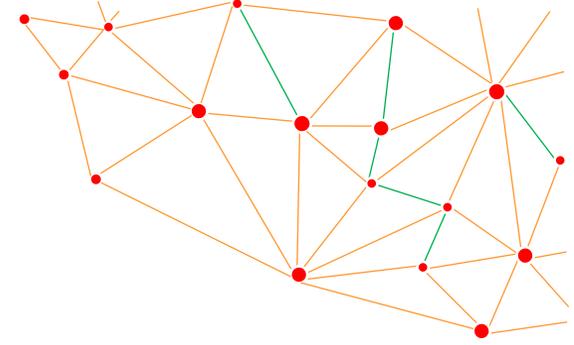
- Стойка (Stoyka) has several representations as a **given name** or a **surname**
- But it also refers to the concepts of:
 - **(body) posture** and
 - **stand**
- BTB-WN contains only one concept - that of **the posture**
- Another problem in this case is that the two pages for these general concepts do not exist, but they are defined only in the disambiguation page. Thus, the annotator has to use a special relation to the disambiguation page

Mid-Observations



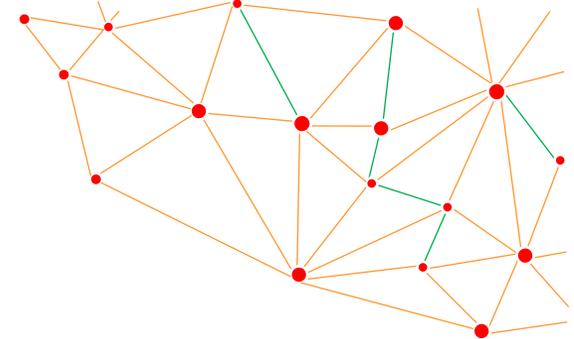
- In more than **78 %** of the cases we establish a correspondence between BTB-WN and the Wikipedia
- Also we have added about **15 %** new concepts and named entities
- The extension would enhance named entity linking, relation extraction and word sense disambiguation
- The main source of enriching BTB-WN appeared to be the NEs and the domain terms as well as MWEs

Limitations



- Wikipedia contains mainly nouns
- For the verbs, adjectives and adverbs other enriching sources should be considered
- Through the derivation relations in WordNet, however, we still could incorporate the presented in Wikipedia deverbal and adjectival nouns
- We envisage to map BTB-WN also to other semantic resources such as Wikidata

Lessons Learnt



- WordNets are not perfect resources since they reflect the grammar of a natural language, a certain conceptualization model of a society and complex lexicalized world knowledge.
- At the same time, WordNets are very necessary resources for a language since they show the hierarchy of interconnected lexical meanings, provide multilingual insights on our cognitive strategies of knowledge organization, and proved to be very useful in a number of NLP tasks.
- WordNets can be viewed as hubs for structuring grammar and semantic knowledge of a language.

Questions Any?

