The CLaDA-BG Dictionary Creation System: Specifics and Perspectives

Zhivko Angelov, Kiril Simov, Petya Osenova, Zara Kancheva
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Sofia, Bulgaria
angelov.zhivko@gmail.com, {kivs,petya,zara}@bultreebank.org

Abstract

The paper reports on the current status of a system for creating dictionaries within the CLaDA-BG infrastructure. The system is called CLaDA-BG-Dict. At the heart of the system lies the lexical thesaurus BTB-Wordnet around which all other language resources for Bulgarian are organized. These are various types of dictionaries (morphological, explanatory, terminological, etc.), ontologies (such as DBpedia), corpora (in-house and external). The specific features and functionalities of the system are discussed with respect to the language resourse integrity. Also, the rationale behind the construction of such a system are given together with an outline of its utility for a number of NLP tasks and for various types of users. The ideas presented as well as the system itself are scalable to integrating resources also for other languages.

1 Introduction

In this paper we present the main principles and perspectives behind the CLaDA-BG Dictionary Creation System — **CLaDA-BG-Dict**. The ultimate goal of its implementation is to support the compilation of new dictionaries by individuals or collaborators with respect to a certain task and through the usage of all the available resources within the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH — CLaDA-BG[1].

We aim to provide a system that supports the whole cycle of creating various types of dictionaries. At the heart of this system lies the Bulgarian BulTreeBank WordNet (BTB-WN) — (Osenova and Simov, 2018). It has been developed as an aggregator of semantic knowledge around which other dictionaries and sources of information (including grammatical, encyclopedic, etc.) have been organized in the form of a(n) (inter)linked knowledge network.

The motivation for the development of the CLaDA-BG-Dict refers to the need for: better control on the consistency in the creation of lexical language resources; user friendly and communicative collaborative environment; better connections among the available resources. Also in the light of open data we expect that there will be more lexicographical data available for reuse in future. This would facilitate the rapid creation of specialised lexicons as well as their publishing and focused usage.

The incentive for the design and implementation of CLaDA-BG-Dict system was the development of BTB-WN. On the one hand, we were aware that there already exist software systems for the creation of other wordnets such as BulNet, GermaNet and Polish Wordnet (plWordNet). However, these systems inevitably reflect the approaches of the creators of these wordnets and thus, they do not support all the functions, needed for the work on BTB-WN. These are: extension of lexical entries structure; mapping to other resources (inflectional lexicons, explanatory dictionaries, bilingual dictionaries, Wikipedia pages, etc.); concordance for selection of examples; ticketing system for identifying and handling errors of various types and history of changes. On the other hand, the workflow on a contemporary wordnet requires

[1] https://clada-bg.eu/en/

Zhivko Angelov, Kiril Simov, Petya Osenova and Zara Kancheva 2023. The CLaDA-BG Dictionary Creation System: Specifics and Perspectives. *Selected papers from the CLARIN Annual Conference 2022. Ed. by Tomaž Erjavec and Maria Eskevich.* Linköping Electronic Conference Proceedings 198, pp. 12–22. DOI: https://doi.org/10.3384/ecp198002

the addition of linguistic information beyond synsets, lemmas, definitions and relations. Such information includes: links to grammatical paradigms, word valencies, links to Wikipedia, mappings to other wordnets, appropriate examples, etc.

For all the reasons presented above, in this paper we present our customized solution for a resource integrating tool. The idea and implementation are scalable also to resources in other languages.

The structure of the paper is as follows: the next section provides a focused review of related works. Section 3 discusses the specifics and functionalities of **CLaDA-BG-Dict** system. Section 4 outlines the language resources that support dictionary creation. Section 5 concludes the paper and presents plans for future work.

2 Related Work

In our work we follow the approaches described in two existing wordnet editing systems —(Henrich and Hinrichs, 2010) and (Naskret et al., 2018). Similarly to Henrich and Hinrichs (2010) we needed to switch from a tool that supported only local editing where synsets were considered within a very limited context to a tool that supports editing of the wordnet data within a larger context. Comparably to both systems we switched from a file oriented presentation of data to a centralized database used via web to support simultaneous work of a team of experts. The most important benefit of this switch is that each member of the working team started to observe the changes made in content and structure at the time they had been made. In addition, the user of the system has the possibility to consult the resources in the database in any moment when this is necessary. Thus, the users have at their disposal a global view over the wordnet. As a consequence, when editing, they might take into account all the data in a connected way instead of partial or isolated views.

Here the following question might arise: Why to develop yet another system when there are already so many out there? We decided to implement our own system because in addition to developing our wordnet, we wanted to support and connect all the language resources we already had incorporated within it. These are: a spelling and grammar dictionary, an explanatory dictionary, bilingual dictionaries, related corpora for providing adequate examples that register various characteristics of the respective meaning. For us the mapping between the existing language resources is set as an important goal. Thus, we wanted to support such mappings as early in the process of the wordnet creation as possible.

Our aim is to extend the current system further towards a full-fledged dictionary writing system. It is envisaged to provide the necessary environment and services for the compilation of ad hoc and task-oriented dictionaries through the access to all the language data – starting from the existing dictionaries, corpora, encyclopedic knowledge, and others.

We are aware that many efforts have already been invested in dictionary creation systems from various points of view: formats and standards; approaches in the representation of the linguistic knowledge; implementation strategies, etc. Here we mention only some of the related work. One of the most influential ongoing frameworks is ELEXIS[2]. After having performed an in-depth survey on the needs of lexicographers[3] — (Kallas et al., 2019), the team behind ELEXIS (p. 62) envisaged 'two complementary sets of tools will be provided: lexicographic workflow tools and crowdsourcing and gamification tools. The first will include a user friendly open-source online dictionary writing system, with the aim to provide the central dictionary writing platform for new lexicography which also includes new possibilities of online collaboration. The other will provide tools for new techniques of dictionary creation, such as explicit or implicit crowdsourcing (gamification).' There are two tools for dictionary creation provided by ELEXIS – OneClick Dictionary[4] and Lexonomy[5] — (Měchura, 2017). The OneClick Dictionary is a dictionary drafting module, a feature of Sketch Engine[6] which produces a machine generated dictionary draft that is later edited by lexicographers in the Lexonomy module. Functionalities such as wordlists,

[2] https://elex.is/

[3] https://elex.is/wp-content/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf

[4] https://github.com/elexis-eu/ocd

[5] https://lexonomy.elex.is/

[6] https://www.sketchengine.eu/

corpora, concordance, thesauri, etc. are also integrated in the tools. Similarly to the CLaDA-BG-Dict system, these tools are applicable also for the tasks of creating glossaries and domain-specific wordlists and dictionaries. Unfortunately, when we started the implementation of CLaDA-BG-Dict system, these tools were not available for a public use. Thus, we plan to customize and adapt them to our framework as much as possible in our future work.

Our current system supports Lexical Markup Framework (LMF) formats but not in its full capacity. LMF files can be uploaded, edited and then saved outside the system. However, not everything from LMF is supported. There is no converter from the internal files into Lemon Standard and back[7]. At this point we rely only on the LMF-based converters. It should be noted once again that we aim to facilitate the work not only of the professional lexicographers but also of any other researcher groups and common users. Thus, we imagine helping teachers to compile a dictionary of minimum words/senses, etc. for their class; or a student to construct incrementally a learner lexicon of Bulgarian related to a language that they know, etc. Within DARIAH-ERIC a standard for representation of dictionaries has been developed — TEI-Lex0.[8] This standard is supported also by ELEXIS. Since it has already been established as a best practice, we plan to use it as well. At the moment we support only the minimum to exchange data in BTB-WN whereas the complete set of import and export formats needs to be implemented. The main focus in implementing the system was put initially on the availability and integration of the resources. Thus, our efforts on ensuring adequate exchanging formats and adherence to the common standards come next.

Last but not least, one of the CLARIN-ERIC Resource Families are Lexica. They are 89 and most of them are monolingual.[9] They are of various types – inflectional, morphological, valency, multiword, stopwords, sentiment, etc. Thus, they are a good source for insights in adding more types of resources and more types of analyses into the system.

3 System Specifics and Functionalities

Initially, CLaDA-BG-Dict was designed and implemented to support the verification and extension of BTB-WN. The motivation for this was that the existing version of BTB-WN was initiated in an XML format within the CLaRK System[10] — (Simov et al., 2001). The XML format used during the creation of earlier versions of BTB-WN was not a standard one. It was designed to facilitate the editing of lexical entries for each synset. Also this XML format had to reflect the incorporation of non-standardised data such as links to Open English Wordnet (OEW), Bulgarian Wikipedia, and others. However, the creation of BTB-WN in this way revealed some shortcomings. As mentioned above, the main problem with working in CLaRK System was that the users had only a local view over the existing Bulgarian synsets because the data with BTB-WN were stored in several XML files, and searches had to be performed within each of them (or in some of them). For instance, it was not easy to observe all the synsets in which a given lemma participates, because they could be in different XML files. Thus, one of the main design solutions was to support the mapping to the OEW with the idea to enhance the multilingual applications and the transfer of information from OEW to BTB-WN. In addition, we needed some system support for the better integration of BTB-WN with other language and knowledge resources for Bulgarian.

The system is a client-server web-based editor using a thick client model. The thick client is installed on the user computer (desktop or laptop). The thick client as a user interface to the system provides a better flexibility with respect to implementing the necessary functionality. It especially facilitates the way to compose several actions during the creation and editing of new synsets, assigning shortcuts and others.

3.1 Initial Acquaintance with the System

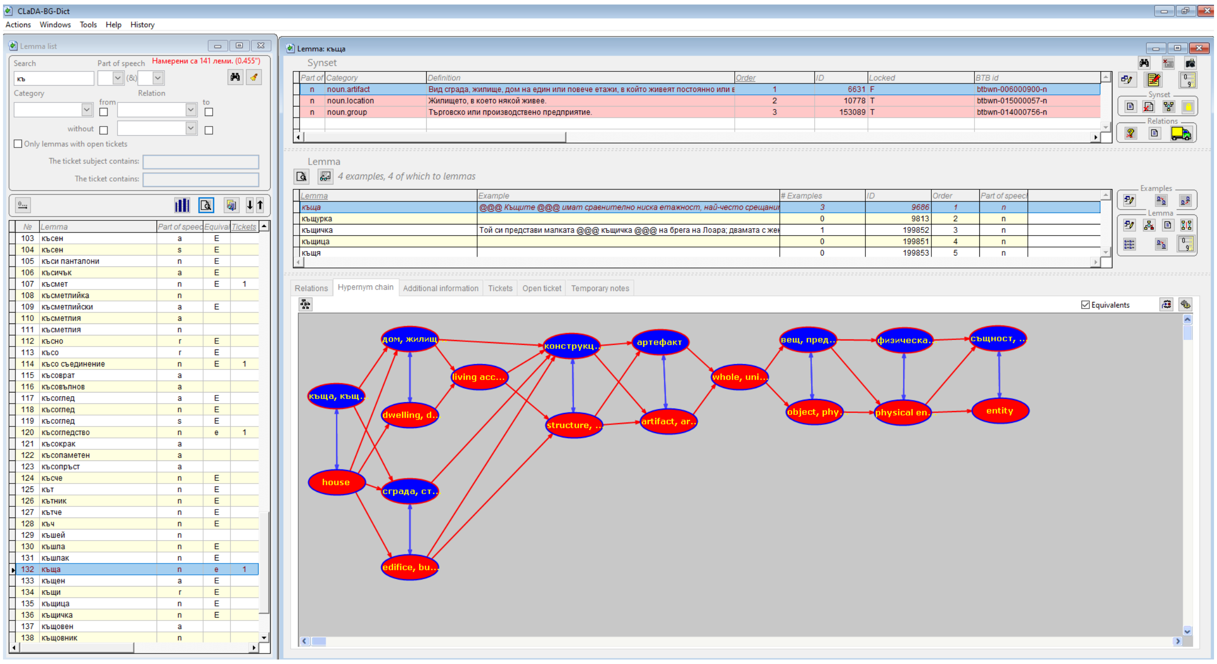
The database is installed on a server and it is accessed online via the web. A relational database is used for storing the data. Two people are not allowed to work on the same synset at the same time. They can

[7] https://www.w3.org/2019/09/lexicog/

[8] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[9] https://www.clarin.eu/resource-families/lexical-resources-lexica

[10] http://bultreebank.org/en/clark/



№	Lemma	Part of speed	Equival	Tickets
103	късен	a	E	
104	късен	s	E	
105	къси панталони	n	E	
106	късичък	a	E	
107	късмет	n	E	1
108	късметлийка	n		
109	късметлийски	a	E	
110	късметлия	a		
111	късметлия	n		
112	късно	r	E	
113	късо	r	E	
114	късо съединение	n	E	1
115	късоврат	a		
116	късовълнов	a		
117	късоглед	a	E	
118	късоглед	n	E	
119	късоглед	s	E	
120	късогледство	n	e	1
121	късокрак	a		
122	късопаметен	a		
123	късопръст	a		
124	късче	n	E	
125	кът	n	E	
126	кътник	n	E	
127	кътче	n	E	
128	къч	n	E	
129	кьшей	n		
130	кьшла	n	E	
131	кьшлак	n	E	
132	къща	n	e	1
133	къщен	a	E	
134	къщи	r	E	
135	къщица	n	E	
136	къщичка	n	E	
137	къщовен	a		
138	къщовник	n		

Part of	Category	Definition	Order	ID	Locked	BTB id
n	noun.artifact	Вид сграда, жилище, дом на един или повече етажи, в който живеят постоянно или е	1	6631	F	btbwn-006000900-n
n	noun.location	Жилището, в което някой живее.	2	10778	T	btbwn-015000057-n
n	noun.group	Търговско или производствено предприятие.	3	153089	T	btbwn-014000756-n

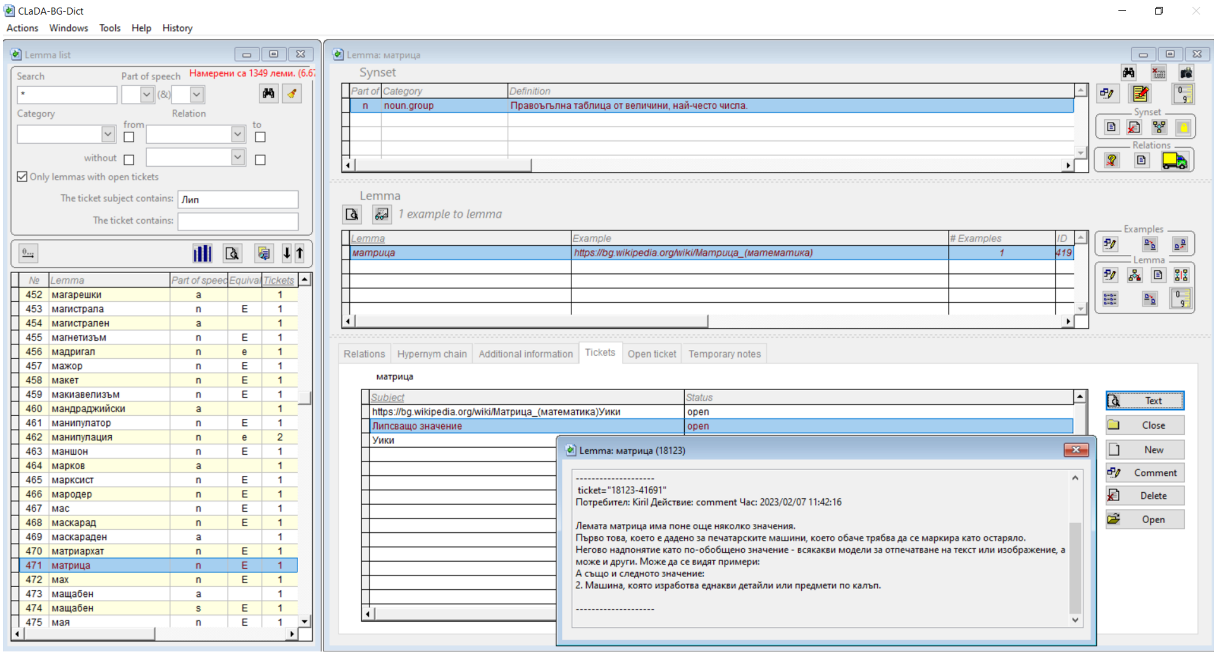
Lemma	Example	# Examples	ID	Order	Part of speec
къща	@@@ Къщите @@@ имат сравнително ниска етажност, най-често срещани	3	9686	1	n
къщурка		0	9813	2	n
къщичка	Той си представи малката @@@ къщичка @@@ на брега на Лоара; двамата с жен	1	199852	3	n
къщица		0	199851	4	n
къщя		0	199853	5	n

Figure 1: A screenshot of the user interface of CLaDA-BG-Dict.

work subsequently on the same synset or simultaneously on different synsets. Also, the system stores in a log file the following information: each editing step, the name of the person who edited, and at what time the edit was done. In this way we might track back states of data and repair errors if necessary. The system reflects our approach towards the next wordnet developments as well as towards the integration of various language resources within any dictionary compilations. This approach reflects the lexicon-grammar interface in a better way. In Fig. 1 a screenshot is presented, which shows a search over lemmas starting with "ка" — this search string is not a word in Bulgarian and only serves as a query which selects a range of lemmas within BTB-WN starting with it. The search string and the result from the search are displayed within the left element of the window in the figure, named *Lemma list*. This part of the window is separated in three parts. The upper part supports searches in the database. Searches can be performed by several criteria: by string, by POS or category, by relations or by the type of the ticket that was assigned to a synset. The list of results from the search is presented in the bottom part of the left element of the window. Each row in this table contains a lemma and the POS of the lemma. It gives information whether there is an equivalent synset in OEW, the number of tickets it has, etc. The search with the string "ка" returns more than 140 lemmas, among which "къща" *kashta* ('house'). The middle part contains icons for the possible operations over lemmas in the list like - sorting, statistics and opening of a selected lemma within an editor form.

Thus, when the lemma къща is selected (as in the figure) and the editor form is opened (displayed on the right side of the window), it can be seen that there are three meanings (synsets) with the categories `noun.artifact`, `noun.location`, and `noun.group` which contain this lemma. The editor form is related to a given lemma (marked in the left upper corner — Lemma: къща). In this way the system allows for the simultaneous opening of several editor forms. These might be used to support the comparison of different synsets for different lemmas. Also they might facilitate the creation of relations between various synsets.

Each editor form consists of three areas: *Synset* area, located at the top part of the form, *Lemma* area, located in the middle part of the form, and *Miscellaneous* area at the bottom of the form. The *Synset area* contains information for the synsets that include the lemma related to the editor form. The information of a given synset includes: the category of the synset; the definition; the order of synsets for the lemma, the



Намерени са 1349 леми. (6.67

Search: *

☐ from ☐ to ☐ without ☐

☑ Only lemmas with open tickets

The ticket subject contains: Лип

The ticket contains:

№	Lemma	Part of speec	Equival	Tickets
452	магарешки	a		1
453	магистрала	n	E	1
454	магистрален	a		1
455	магнетизъм	n	E	1
456	мадригал	n	e	1
457	мажор	n	E	1
458	макет	n	E	1
459	макиавелизъм	n	E	1
460	мандраджийски	a		1
461	манипулатор	n	E	1
462	манипулация	n	e	2
463	маншон	n	E	1
464	марков	a		1
465	марксист	n	E	1
466	мародер	n	E	1
467	мас	n	E	1
468	маскарад	n	E	1
469	маскараден	a		1
470	матриархат	n	E	1
471	матрица	n	E	1
472	мах	n	E	1
473	мащабен	a		1
474	мащабен	s	E	1
475	мая	n	E	1

Part of	Category	Definition
n	noun.group	Правоъгълна таблица от величини, най-често числа.

1 example to lemma

Lemma	Example	# Examples	ID
матрица	https://bg.wikipedia.org/wiki/Матрица_(математика)	1	419

матрица

Subject	Status
https://bg.wikipedia.org/wiki/Матрица_(математика)Уики	open
Липсващо значение	open
Уики	

Lemma: матрица (18123)

```
--------------------
 ticket="18123-41691"
Потребител: Kiril Действие: comment Час: 2023/02/07 11:42:16

Лемата матрица има поне още няколко значения.
Първо това, което е дадено за печатарските машини, което обаче трябва да се маркира като остаряло.
Негово надпонятие като по-обобщено значение - всякакви модели за отпечатване на текст или изображение, а
може и други. Може да се видят примери:
А също и следното значение:
2. Машина, която изработва еднакви детайли или предмети по калъп.

--------------------
```

Figure 2: A screenshot of the tickets in CLaDA-BG-Dict.

internal ID of the synset (unique and unchangeable within the database), the locking information, and the BTB-WN ID of the synset. Next to the table are the operations that can be performed over a given synset. These include: editing of the definition (if not locked), export of the synset in a textual form, reordering of the synsets of a lemma (see below), creation of a new synset, deletion of a synset, manipulation of relations. The *Lemma area* consists also of two elements – a table and a tool pane. In the table a list of lemmas in the synset is presented. Each lemma is associated with some examples for the given synset, its paradigm and internal ID. The lemmas in the same synset are ordered with respect to how well they represent the meaning of the synset. The lemma tools include manipulation of examples, editing of lemmas, editing of the paradigms, and ordering of the lemmas. The *Miscellaneous area* consists of several tabs presenting different parts of information like relations for the synset, hierarchy of synsets, information from associated machine readable dictionaries, tickets, and temporal notes on some information in the synsets. In the above figure – in the Miscellaneous area – a graphic representation of a hierarchy of a noun is given and also the mapping to the English synsets. The graphic of relations shows the hierarchy of the first synset – noun. artifact. It is a hyponym of *building*, *construction*, *artefact*, and on the highest level – of *physical entity* and *entity*.

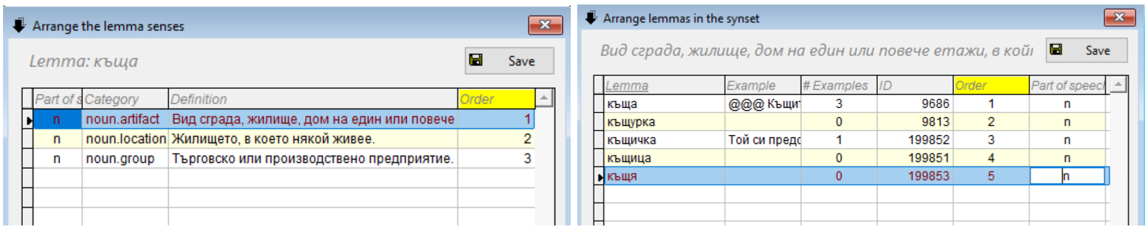
Regarding the BTB-WN, the system provides information about a selected lemma: its meanings (synsets) and associated examples; its internal relations as well as the mappings to the OEW; it also provides the ratio among the used relations. In case of equivalent synsets between BTB-WN and OEW, the Bulgarian synset inherits all the relations from the English synsets. In cases when these equivalent synsets have also corresponding hypernyms or hyponyms, the inherited relations enrich them as well.[11] After the relations have been inherited, the users have the possibility to change them — delete some of them when not applicable or add new ones. The system also supports definitions of new relations and some (limited) inference with reverse and transitive operations. In addition, domain and range restrictions are taken into account.

The system is equipped with a ticket module. Thus, the workflow is organized in a more structured way with respect to the various expertise and responsibilities. Lemmas can be marked in a certain way that

[11] Given that this transfer of relations is correct for the concepts represented by the synsets.

calls for the intervention of a more experienced user. Thus, a user could assign a ticket to a lemma in case of identification of some error, or suggest an edit of a particular synset. Each ticket contains two parts: a textual description of the problem, and a related topic (subject). There is a list of predefined topics for the tickets, created with respect to the workflow on BTB-WN, but other types might be added whenever needed. The list of current ticket subjects includes: *Edit synset*, *Missing sense*, *Wrong hypernym*, *Part of speech*, *General remark*, *Wrong equivalent*, *Discussion*, *Missing relation*, and *Link to Wikipedia*. In Fig. 2 an example is given. In this case the ticket relates to some missing senses for the selected lemma and provides suggestions for these senses. A more experienced user checks all the lemmas with such ticket subjects and approves the existing suggestions or adds the appropriate senses. The system allows for the users to search for lemmas with certain types of tickets. A result from such a search is given in left part of the figure. The search is for all lemmas that are marked with the subject for missing senses. The *Link to Wikipedia* subject is currently used for adding a Wikipedia link to the corresponding lemma in one of its senses. The tickets might be commented by other users (depending on their editing rights), then resolved and deleted. Information is also available about the author of the ticket, the date and the time when it was created, commented or deleted, who and when processed it.

Lemmas and synsets within the wordnets are in many-to-many relations which means that a lemma could belong to several synsets whereas a given synset could have more than one lemma. In both cases the lemmas and the synsets are not equal in their usages. Thus it is important to rank them with respect to their relevance. Currently, we rely only on the users' intuition for this step. Thus we aim at some initial lemma ordering for each synset with more than one lemma. In some cases the ordering is performed automatically. Such cases include, for example, the synsets for professions where the masculine form precedes the feminine one. Also in case of verbal synsets with imperfective and perfective aspect lemmas, the imperfective one comes first. The user interfaces for the ranking are given in Fig. 3. On the left side, the dialog for the arrangement of synsets is given. The user has a granted access to the categories and definitions of each synset. Thus, they can make an informed decision about the ranking. On the right, a similar dialog is presented where the user can check the synsets whose lemmas need to be ordered. In both cases the user could modify the ranking values. These rankings are used in different applications as Word Sense Disambiguation, selection of appropriate lemmas in text generation, etc. Needless to say, the manual ranking is subjective and thus not reliable enough, so in future more information about relations between lemmas and senses will be added to BTB-WN.



Lemma: къща

Part of s	Category	Definition	Order
n	noun.artifact	Вид сграда, жилище, дом на един или повече	1
n	noun.location	Жилището, в което някой живее.	2
n	noun.group	Търговско или производствено предприятие.	3

Вид сграда, жилище, дом на един или повече етажи, в кой

Lemma	Example	# Examples	ID	Order	Part of speec
къща	@@@ Къщи	3	9686	1	n
къщурка		0	9813	2	n
къщичка	Той си предс	1	199852	3	n
къщица		0	199851	4	n
къщя		0	199853	5	n

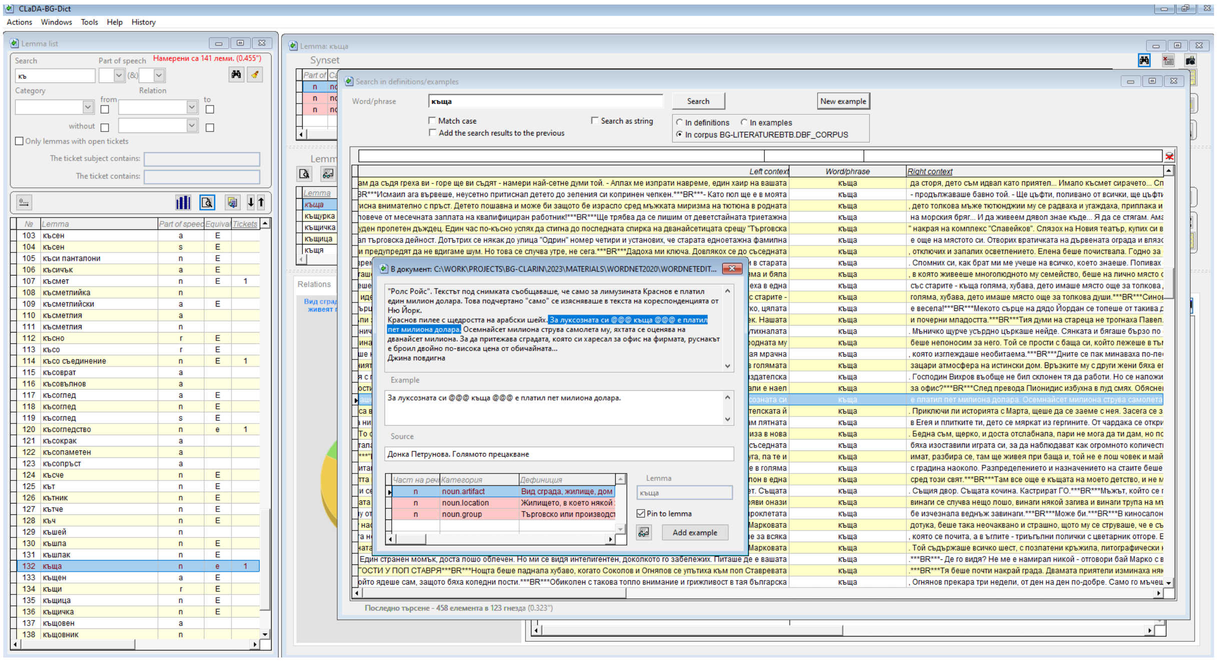
Figure 3: Arranging synsets in CLaDA-BG-Dict.

Our system provides more functionalities than the ones, introduced above. One of them is the access to corpora and machine readable dictionaries. It will be discussed in more detail in the next section. The remaining functionalities solve some smaller tasks like shortcuts assignments, procedures for automatic relation addition and similar.

3.2 Integration of Corpora and Other Dictionaries

In this section two of the main sources of information used by the lexicographers in their work — corpora and dictionaries — are discussed with respect to their integration within the CLaDA-BG-Dict System. The corpora are mainly used for searching examples for the various lemma senses and for finding new or missing senses. The dictionaries are valuable sources of different kinds of lexical knowledge.

The system allows for searching and adding example sentences directly (see Fig. 4). Users can provide any corpora relevant to their work. When selecting an example, the user can pin it to the corresponding



Search: къ Part of speech (&) Намерени са 141 леми. (0.455")
Category from □ Relation to □
without □ □
□ Only lemmas with open tickets
The ticket subject contains:
The ticket contains:

№	Lemma	Part of speech	Equival	Tickets
103	късен	a	E	
104	късен	s	E	
105	къси панталони	n	E	
106	късичък	a	E	
107	късмет	n	E	1
108	късметлийка	n		
109	късметлийски	a	E	
110	късметлия	a		
111	късметлия	n		
112	късно	r	E	
113	късо	r	E	
114	късо съединение	n	E	1
115	късоврат	a		
116	късовълнов	a		
117	късоглед	a	E	
118	късоглед	n	E	
119	късоглед	s	E	
120	късогледство	n	e	1
121	късокрак	a		
122	късопаметен	a		
123	късопръст	a		
124	късче	n	E	
125	кът	n	E	
126	кътник	n	E	
127	кътче	n	E	
128	кьч	n	E	
129	кьшей	n		
130	кьшла	n	E	
131	кьшлак	n	E	
132	къща	n	e	1
133	къщен	a	E	
134	къщи	r	E	
135	къщица	n	E	
136	къщичка	n	E	
137	къщовен	a		
138	къщовник	n		

Search in definitions/examples

Word/phrase: къща Search New example
□ Match case □ Search as string
□ Add the search results to the previous
○ In definitions ○ In examples
● In corpus BG-LITERATUREBTB.DBF_CORPUS

Left context	Word/phrase	Right context
ам да съдя греха ви - горе ще ви съдят - намери най-сетне думи той. - Аллах ме изпрати навреме, един хаир на вашата	къща	да сторя, дето съм идвал като приятел... Имало късмет сирачето... Сп
BR***Исмаил ага вървеше, неусетно притиснал детето до зеления си копринен чепкен.***BR***- Като гюл ще е в моята	къща	- продължаваше бавно той. - Ще цъфти, поливано от всички, ще цъфти
исна внимателно с пръст. Детето пошавна и може би защото бе израсло сред мъжката миризма на тютюна в родната	къща	, дето толкова мъже тютюнджии му се радваха и угаждаха, приплака и
повече от месечната заплата на квалифициран работник!***BR***Ще трябва да се лишим от деветстайната триетажна	къща	на морския бряг... И да живеем дявол знае къде... Я да се стягам. Ама
уден пролетен дъждец. Един час по-късно успях да стигна до последната спирка на дванайсетицата срещу "Търговска	къща	" накрая на комплекс "Славейков". Слязох на Новия театър, купих си в
ал търговска дейност. Дотътрих се някак до улица "Одрин" номер четири и установих, че старата едноетажна фамилна	къща	е още на мястото си. Отворих вратичката на дървената ограда и влязо
и предупредят да не вдигаме шум. Но това се случва утре, не сега.***BR***Дадоха ми ключа. Довлякох се до съседната	къща	, отключих и запалих осветлението. Елена беше почиствала. Годно за
и в старата	къща	. Спомних си, как брат ми ме учеше на всичко, което знаеше. Попивах
ма и бяла	къща	, в която живееше многолюдното му семейство, беше на лично място
еха в една	къща	със старите - къща голяма, хубава, дето имаше място още за толкова ,
с старите -	къща	голяма, хубава, дето имаше място още за толкова души.***BR***Синов
ко, цялата	къща	е весела!***BR***Мекото сърце на дядо Йордан се топеше от такива
ек. Нашата	къща	и почерни младостта.***BR***Тия думи на стареца не трогнаха Павел.
утихналата	къща	. Мъничко щурче усърдно църкаше нейде. Сянката и бягаше бързо по
родната му	къща	беше непоносим за него. Той се прости с баща си, който лежеше в тъ
ая мрачна	къща	, която изглеждаше необитаема.***BR***Дните се пак минаваха по-ле
в голямата	къща	зацари атмосфера на истински дом. Връзките му с други жени бяха е
здателска	къща	. Господин Вихров въобще не бил склонен тя да работи. Но се наложи
али е наел	къща	за офис?***BR***След превода Пионидис избухна в луд смях. Обясне
созната си	къща	е платил пет милиона долара. Осемнайсет милиона струва самолета
телската й	къща	. Приключи ли историята с Марта, щеше да се заеме с нея. Засега се з
м лятната	къща	в Егея и плитките ти, дето се мяркат из гергините. От чардака се откри
иза в нова	къща	. Бедна съм, щерко, и доста отслабнала, пари не мога да ти дам, но по
съседната	къща	бяха изоставили играта си, за да наблюдават как огромното количест
уга, па те и	къща	имат, разбира се, там ще живея при баща и, той не е лош човек и май
е в голяма	къща	с градина наоколо. Разпределението и назначението на стаите беше
лон в една	къща	сред този свят.***BR***Там все още е къщата на моето детство, и не м
ет. Същата	къща	. Същия двор. Същата кочина. Кастрират ГО.***BR***Мъжът, който се
ряви онази	къща	винаги се случва нещо лошо, винаги някой загива и винаги трупа на м
проклетата	къща	бе изчезнала веднъж завинаги.***BR***Може би.***BR***В киносалон
Марковата	къща	дотука, беше така неочаквано и страшно, щото му се струваше, че е с
е за всяка	къща	, която се почита, а в ъглите - триъгълни полички с цветарник отгоре. Е
Марковата	къща	. Той съдържаше всичко шест, с позлатени кръжила, литографически
Един странен момък, доста лошо облечен. Но ми се видя интелигентен, доколкото го забележих. Питаше де е вашата	къща	.***BR***- Де го видя? Не ме е намирал никой - отговори бай Марко с в
ГОСТИ У ПОП СТАВРЯ***BR***Нощта беше паднала хубаво, когато Соколов и Огнянов се упътиха към поп Ставревата	къща	.***BR***Тя беше почти накрай града. Двамата приятели изминаха ня
ойто ядеше сам, защото бяха коледни пости.***BR***Обиколен с такова топло внимание и грижливост в тая българска	къща	, Огнянов прекара три недели, от ден на ден по-добре. Само го мъчец

Последно търсене - 458 елемента в 123 гнезда (0.323")

В документ: C:\WORK\PROJECTS\BG-CLARIN\2023\MATERIALS\WORDNET2020\WORDNETEDIT...

"Ролс Ройс". Текстът под снимката съобщаваше, че само за лимузината Краснов е платил един милион долара. Това подчертано "само" се изясняваше в текста на кореспонденцията от Ню Йорк.
Краснов пилее с щедростта на арабски шейх. За луксозната си @@@ къща @@@ е платил пет милиона долара. Осемнайсет милиона струва самолета му, яхтата се оценява на дванайсет милиона. За да притежава сградата, която си харесал за офис на фирмата, руснакът е броил двойно по-висока цена от обичайната...
Джина повдигна

Example: За луксозната си @@@ къща @@@ е платил пет милиона долара.

Source: Донка Петрунова. Голямото прецакване

Част на речта	Категория	Дефиниция
n	noun.artifact	Вид сграда, жилище, дом
n	noun.location	Жилището, в което някой
n	noun.group	Търговско или производст

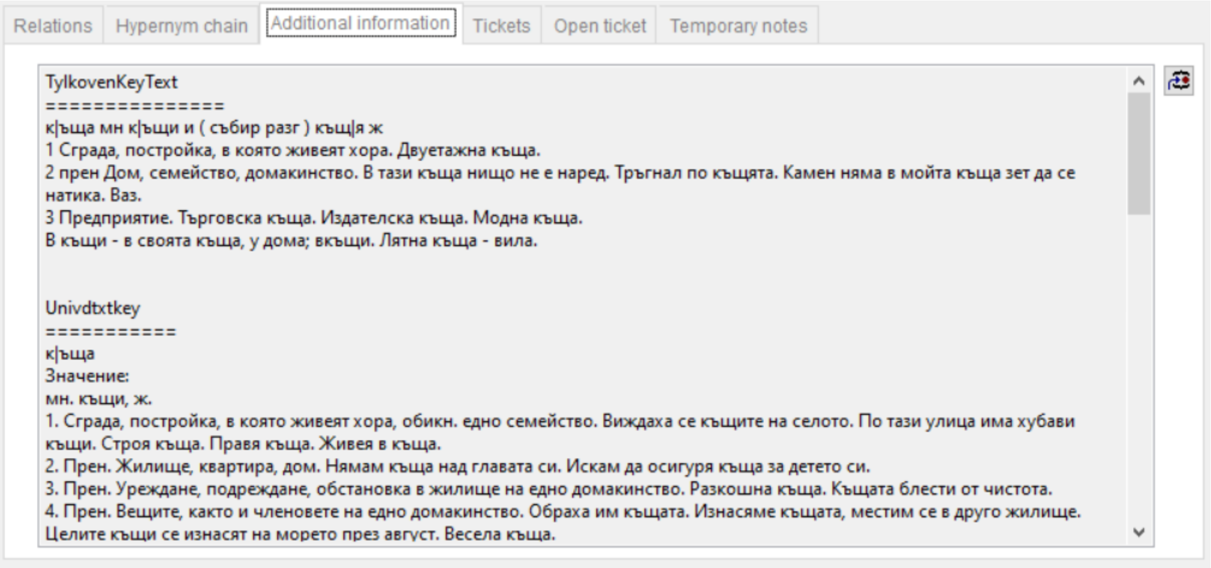
Lemma: къща
☑ Pin to lemma
Add example

Figure 4: Search for examples in a corpus integrated in CLaDA-BG-Dict.

synset and lemma while the source of the example is copied automatically. The search could be performed by lemmas or strings. In future, the search engine will support also regular expressions. The system suggests that sentences are included as examples, but if the users decide otherwise, they can include also bigger contexts from source texts. The latter is especially important for the selection of examples that do not allow ambiguous interpretations. The assignment of examples to lemmas and synsets resembles the process of text annotation with senses from BTB-WN. For that reason, there are two special cases of corpora in the system that were already assigned to lemmas and synsets: the set of definitions and the set of examples. The user has the possibility to search in them and in this way to annotate all the words within definitions and examples. The intended usage of this option is the creation of corpora annotated with senses from BTB-WN, similarly to SEMCOR (Miller et al., 1993) and Gloss corpus (Rademaker et al., 2019). In addition to being used as sources for assigning examples to the existing synsets, the corpora are extensively used also for detecting new senses that are neither in the current version of BTB-WN, nor in the dictionaries. Currently we can only rely on the sorting functionality of the concordance with respect to the found items and their contexts. After having been sorted, the examples are checked one by one. In future we plan to use similarity measures over the context in order to cluster the concordance lines.

In addition to corpora we consider the access to existing machine readable dictionaries within the CLaDA-BG-Dict System as an important resource to be consulted during the creation of BTB-WN. The system provides access to four electronic dictionaries which are aligned through the lemmas they share. When an entry contains information about several lemmas, it is aligned to the other dictionaries through each of these lemmas. In this way the information for a given lemma is accessible through each of the lemmas in any of these dictionaries. Thus, users could observe all the information coming from the various dictionaries simultaneously. The four dictionaries integrated and actively used in the CLaDA-BG-Dict include: one explanatory dictionary — (Popov, 1994), one inflectional dictionary of Bulgarian — (Popov et al., 1998), (Popov et al., 2003), and two Bulgarian-to-English dictionaries — one freely available on the web and one based on our own Bulgarian vocabulary for bilingual dictionaries. The bilingual dictionaries are particularly useful for the selection of appropriate English equivalents from the EOW as well as for providing information about the number of senses for a given lemma. The explanatory dictionary includes also information about idioms with the selected lemma, so they can be used as a source for creating synsets with multiword expressions. A problem which occurs during the integration of the various dictionaries is that there could be discrepancies among them of various kinds. For example, the dictionaries in the system provide as a rule different number of senses for a lemma. The reasons for this could be many but some of them are: some dictionaries include also archaic and dialectal senses and/or tend to distinguish among very similar senses, while others are more general and present only contemporary and/or gross-level senses. In addition, dictionaries are published at different times, so they show the senses typical for two or more different periods. Such contradictions between dictionaries are normal and should be expected, since all of them could follow different approaches and goals. An example of this issue from BTB-WN is the case with the noun чета: its most frequent sense found in all dictionaries is *a group of rebellions in liberation struggles*, but one of the integrated in the CLaDA-BG-Dict system dictionaries includes also an archaic sense *a group for common work* and a rare metaphorical meaning *gymnastics or other sport group*. The newer and more general dictionary in the system presents only the first most frequent sense nowadays. In such case, the lexicographers should follow the corresponding guidelines for the selection of senses.

Another challenge is related to the orthography – as mentioned above, the incorporated dictionaries in CLaDA-BG-Dict are published at different periods, so some them are not complying to the current orthographic rules. At the same time, they all contain valuable information for senses, morphology, etc. and they are worth to be considered. The problem which however comes is that some lemmas would be flagged as not present in BTB-WN just because they are written differently.

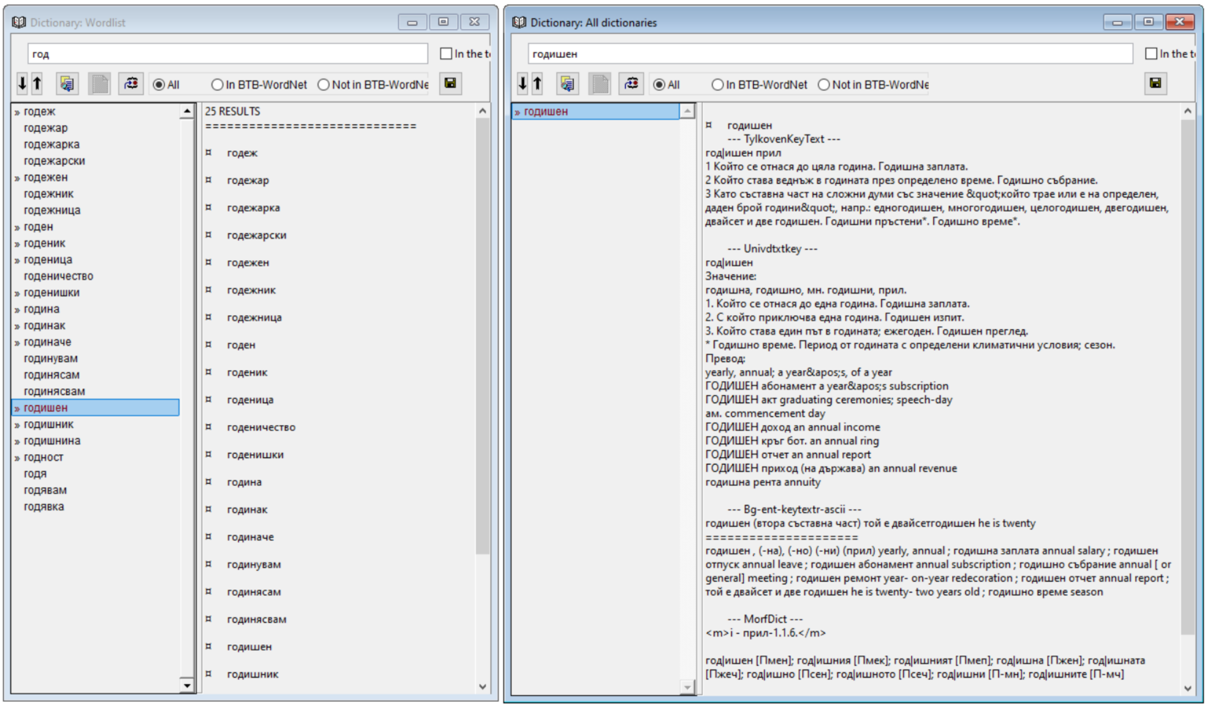


```
TylkovenKeyText
===============
к|ъща мн к|ъщи и ( събир разг ) къщ|я ж
1 Сграда, постройка, в която живеят хора. Двуетажна къща.
2 прен Дом, семейство, домакинство. В тази къща нищо не е наред. Тръгнал по къщята. Камен няма в мойта къща зет да се
натика. Ваз.
3 Предприятие. Търговска къща. Издателска къща. Модна къща.
В къщи - в своята къща, у дома; вкъщи. Лятна къща - вила.


Univdtxtkey
===========
к|ъща
Значение:
мн. къщи, ж.
1. Сграда, постройка, в която живеят хора, обикн. едно семейство. Виждаха се къщите на селото. По тази улица има хубави
къщи. Строя къща. Правя къща. Живея в къща.
2. Прен. Жилище, квартира, дом. Нямам къща над главата си. Искам да осигуря къща за детето си.
3. Прен. Уреждане, подреждане, обстановка в жилище на едно домакинство. Разкошна къща. Къщата блести от чистота.
4. Прен. Вещите, както и членовете на едно домакинство. Обраха им къщата. Изнасяме къщата, местим се в друго жилище.
Целите къщи се изнасят на морето през август. Весела къща.
```

Figure 5: Access to dictionaries within CLaDA-BG-Dict via an editor form.

The access to the dictionaries is provided in two ways. The first one is from the editor form. In Fig. 5 information about the lemma къща in two explanatory dictionaries (which also provide idioms with the given lemma) is observed. The search is by the lemma associated to the editor form. The information from the dictionaries is presented in a tab from the Miscellaneous area. Such a look up in dictionaries is very convenient for quick checks of the various definitions, examples, English corresponding lemmas during the editing of existing synsets in BTB-WN. The second mechanism for a look up is independent from the editor form. There is an option for each dictionary to be opened in its own form, or another option when all dictionaries are opened in one form. The availability of these independent forms allows for searches for arbitrary lemmas, related lemmas,etc. A useful mechanism for access to the dictionaries is through the so-called Wordlist form. In this form a list of lemmas provided by the user is opened. The lemmas are checked whether there are already synsets related to them in the BTB-WN. If not, the search is done by pointing to a lemma in the Wordlist form and using a search shortcut which copies the lemma to the dictionary form to perform the search.



год

» годеж
годежар
годежарка
годежарски
» годежен
годежник
годежница
» годен
» годеник
» годеница
годеничество
» годенишки
» година
» годинак
» годиначе
годинувам
годинясам
годинясвам
» годишен
» годишник
» годишнина
» годност
годя
годявам
годявка

25 RESULTS
==============================

¤ годеж
¤ годежар
¤ годежарка
¤ годежарски
¤ годежен
¤ годежник
¤ годежница
¤ годен
¤ годеник
¤ годеница
¤ годеничество
¤ годенишки
¤ година
¤ годинак
¤ годиначе
¤ годинувам
¤ годинясам
¤ годинясвам
¤ годишен
¤ годишник

годишен

» годишен

¤ годишен
--- TylkovenKeyText ---
год|ишен прил
1 Който се отнася до цяла година. Годишна заплата.
2 Който става веднъж в годината през определено време. Годишно събрание.
3 Като съставна част на сложни думи със значение &quot;който трае или е на определен, даден брой години&quot;, напр.: едногодишен, многогодишен, целогодишен, двегодишен, двайсет и две годишен. Годишни пръстени*. Годишно време*.

--- Univdtxtkey ---
год|ишен
Значение:
годишна, годишно, мн. годишни, прил.
1. Който се отнася до една година. Годишна заплата.
2. С който приключва една година. Годишен изпит.
3. Който става един път в годината; ежегоден. Годишен преглед.
* Годишно време. Период от годината с определени климатични условия; сезон.
Превод:
yearly, annual; a year&apos;s, of a year
ГОДИШЕН абонамент a year&apos;s subscription
ГОДИШЕН акт graduating ceremonies; speech-day
ам. commencement day
ГОДИШЕН доход an annual income
ГОДИШЕН кръг бот. an annual ring
ГОДИШЕН отчет an annual report
ГОДИШЕН приход (на държава) an annual revenue
годишна рента annuity

--- Bg-ent-keytextr-ascii ---
годишен (втора съставна част) той е двайсетгодишен he is twenty
=====================
годишен , (-на), (-но) (-ни) (прил) yearly, annual ; годишна заплата annual salary ; годишен отпуск annual leave ; годишен абонамент annual subscription ; годишно събрание annual [or general] meeting ; годишен ремонт year- on-year redecoration ; годишен отчет annual report ; той е двайсет и две годишен he is twenty- two years old ; годишно време season

--- MorfDict ---
<m>i - прил-1.1.6.</m>

год|ишен [Пмен]; год|ишния [Пмек]; год|ишният [Пмеп]; год|ишна [Пжен]; год|ишната [Пжеч]; год|ишно [Псен]; год|ишното [Псеч]; год|ишни [П-мн]; год|ишните [П-мч]

Figure 6: Look up in the dictionaries within CLaDA-BG-Dict in an independent form and with links to the Wordlist.

In Fig. 6 the two forms — All dictionary and Wordlist, are represented. The search in both forms is through regular expressions. In Wordlist we see all the words matching the search query. Selecting a lemma from the list and searching in dictionaries form would provide information from all lexicons.

With relying on different Wordlists, the users could examine some sets of lemmas in BTB-WN selected by certain criteria. In our work we consider several such sets like vocabularies corresponding to Bulgarian learners' levels like A1-A2, B1-B2, C; vocabulary for secondary school students, etc.

4 Conclusions and Future Work

As it was frequently stressed above, we aim to provide a system where the user will be able to exploit all the available dictionaries, corpora and services. At the same time (s)he will have the possibility to not only statically consult other dictionaries but also to search within corpora, make concordances, establish mappings, save and make publicly available the results of their work.

In our view the necessary minimum of functionalities of such a system would include: an editor of lexical entries that supports different structures of interrelated elements; access points to existing dictionaries and corpora; various types of searches and concordances, etc. BTB-WN has been fully developed in this system and serves as a connector to other dictionaries and corpora through its synset and lemma information.

In addition to uploading and making accessible new dictionaries, the system also supports mappings to Wikipedia via the inclusion of Wikipedia article URIs to the corresponding synsets. For now this operation works for equivalent concepts only, but more elaborated set of relations are necessary.[12]

Each of the included language resources inherits its structure defined in some standard (with some modification if necessary). For example, for a given lexicon included in the system the structure of the lexical entry will be presented in TEI Lex0[13]. Some other lexicon might be presented in Lemon or LMF. Thus, the user will be able to refer to the structure of the various lexicons, to extract parts from

[12] We plan to adopt an already existing schema like SKOS, LEMON, etc.

[13] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

the lexicon entries, to combine elements from different types of lexicons. This will allow easy ways of reusing the available data. Another benefit of the system is that it can keep track on the provenance of the work threads. Currently the system keeps information about each editing operation, but we think more elaborate model is necessary.

When the new dictionary (lexicon) is shared within the system together with the relations to other resources, the result will be a valuable resource not only for supporting the future dictionary creation, but also for automatic processing.

CLaDA-BG-Dict is an editor, which could be used for both tasks – creating lexical databases like wordnets and ccompiling traditional types of dictionaries. But what is more – it provides possibilities of linking the available data in many ways depending on the goal. CLaDA-BG-Dict has already been successfully used for editing of more than 19 000 synsets that were created at earlier stages in an XML format, and for the addition of around 14 000 synsets together with appropriate examples. It thus provides quick access to various types of linguistic resources and information – dictionaries, corpora, concordance, etc. The resources are accessible in the system, so any kind of checks could be performed by the user in the same environment.

Our vision for future is to enhance replicability and re-usage of dictionary compilation for specific purposes as much as possible. In this way we believe that the work of dictionary creators and dictionary users will be facilitated and enriched.

Last but not least, in its beta-version now the system uses its own format for uploading corpora and other digitally-born or digitized dictionaries. However, it is planned that the system conforms to the common standards such as TEI, TEI LEX0, Lemon, etc. All the participating resources will be made available through the CLaDA-BG repository and dedicated web services.

Acknowledgements

This work was supported by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-301/17.12.21.

References

Henrich, V. and Hinrichs, E. 2010. *GernEdiT - The GermaNet Editing Tool.* Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)

Kallas, J. and Koeva, S. and Kosem, I. and Langemets, M. and Tiberius, C. 2019 D1.1. *Lexicographic practices in Europe: a survey of user needs.* ELEXIS - European Lexicographic Infrastructure. H2020 project.

Měchura, M 2017. *Introducing Lexonomy: an open-source dictionary writing and publishing system.* In: Proceedings of eLex 2017 conference. Leiden: Lexical Computing, 2017. p. 662–679.

Miller, g. and Leacock, C. and Tengi, R. and Bunker, R. 1993. *A semantic concordance.* Proceedings of the workshop on Human Language Technology. pages 303-–308. Association for Computational Linguistics.

Naskret, T. and Dziob, A. and Piasecki, M. and Saedi, Ch. and Branco, A. 2018. *WordnetLoom – a Multilingual Wordnet Editing System Focused on Graph-based Presentation*. Proceedings of the 9th Global Wordnet Conference. pp. 190–199

Osenova, P., Simov, K. 2018. The data-driven Bulgarian WordNet: BTBWN. Cognitive Studies – Études cognitives, 2018(18) (2018) `https://doi.org/10.11649/cs.1713`

Popov, D. 1994. *Bulgarian Explanatory Dictionary. Extended and Updated Edition. (in Bulgarian)* Nauka i izkustvo. Sofia, Bulgaria

Popov, D. and Simov, K. and Vidinska, S. 1998. *Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language (in Bulgarian)* Atlantis KL, Sofia, Bulgaria

Popov, D. and Simov, K. and Vidinska, S. and Osenova, P. 2003. *Spelling Dictionary of Bulgarian Language. (in Bulgarian)* Nauka i izkustvo. Sofia, Bulgaria

Rademaker, A. and Cuconato, B. and Cid, A. and Tessarollo, A. and Andrade, H. 2019. *Completing the Princeton Annotated Gloss Corpus Project.* Proceedings of the 10th Global Wordnet Conference, pages 378-–386, Wroclaw, Poland. Global Wordnet Association.

Simov, K. and Peev, Z. and Kouylekov, M. and Simov, A. and Dimitrov, M. and Kiryakov, A. 2001. *CLaRK-an XML-based system for corpora development*. Proceedings of the Corpus Linguistics 2001 Conference. pp. 558–560