The CLaDA-BG Dictionary Creation System: Specifics and Perspectives

Zhivko Angelov, Kiril Simov, Petya Osenova, Zara Kancheva
IICT, BAS

CLARIN Annual Conference
10-12 September 2022

Plan of the Talk

- Introduction
- Related Work
- System Specifics and Functionalities
- Supporting Language Resources
- Conclusions

BulTreeBank CLaDA BG

Introduction: Aim

- We present the main principles and perspectives behind the CLaDA-BG Dictionary Creation System - CLaDA-BG-Dict
- The ultimate goal 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

BulTreeBank CLaDA BG

Introduction: Our Idea

- At the heart of this system lies the Bulgarian BulTreeBank WordNet (BTB-WN)
- 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

Introduction: Our Motivation

- Better control on the consistency in the creation of lexical language resources
- User friendly and communicative collaborative environment
- Better connections among the available resources
- We build on these best practices – systems for construction of other wordnets like GermaNet, Polish Wordnet and BulNet. However, we needed a system that reflects our rhythm of work.
- Functionality to provide links to grammatical paradigms, valences, links to Wikipedia, mappings to other wordnets, appropriate examples, etc.

Related Work

- We follow the approaches described in two existing wordnet editing systems: (Henrich and Hinrichs, 2010) and (Naskret et al., 2018)
- Many efforts have already been invested in dictionary creation systems:
  - ELEXIS - lexicographic workflow tools and crowdsourcing and gamification tools
  - DARIAH-ERIC – dictionary representation standard TEI-Lex0
  - CLARIN-ERIC - Lexica Resource Family overviews 89 lexica

System Specifics and Functionalities: the start

- The existing version of the BTB-WN was initiated in an XML format within the CLaRK System
- However: the lexicographers had only a local view (without an underlying database) over the existing Bulgarian synsets
- Support was needed to mapping the Open English WordNet (OEW)
- Support was needed for the better integration of BTB-WN with other language and knowledge resources for Bulgarian

System Specifics and Functionalities: the technicalities

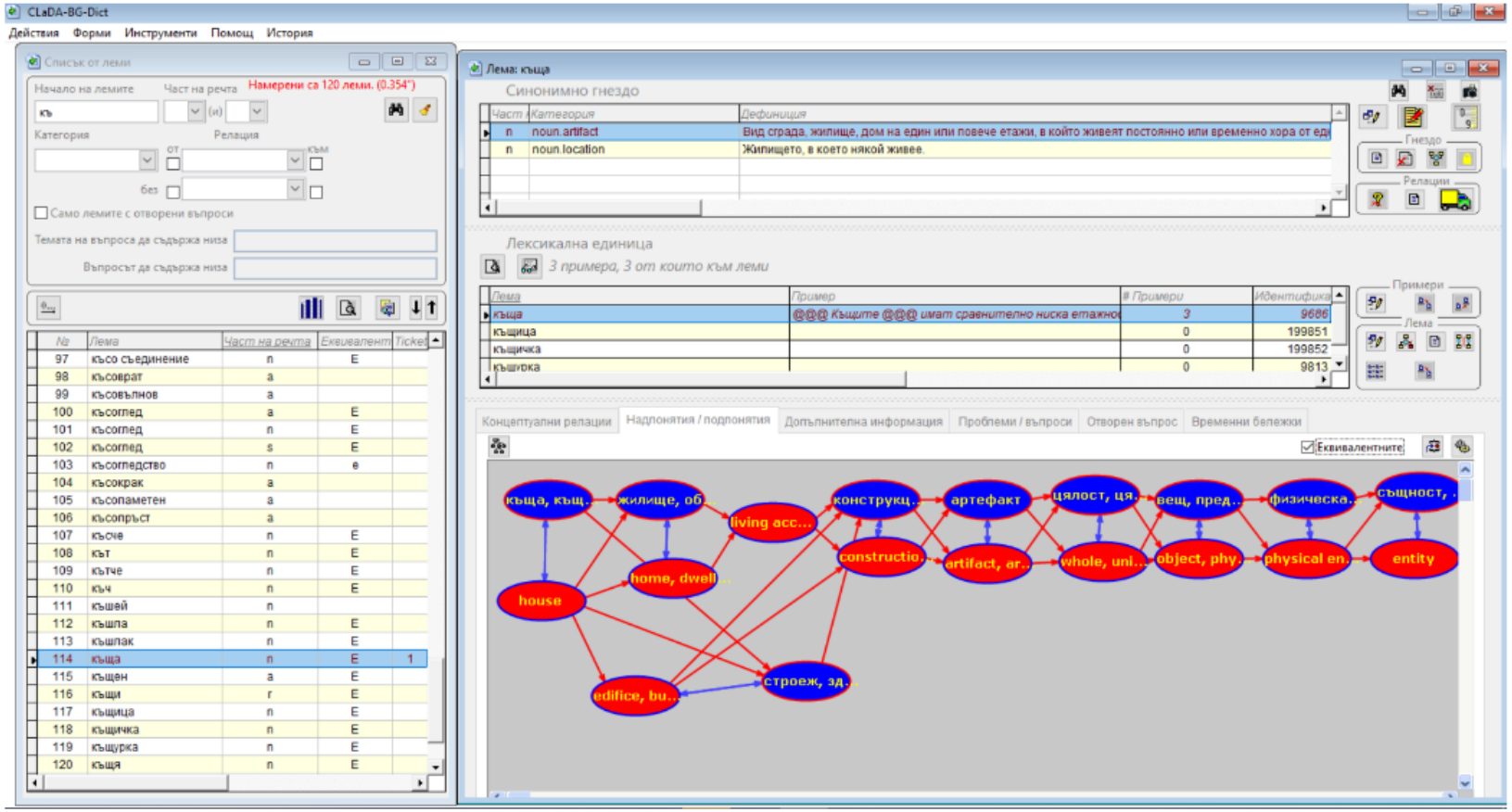
- The system is a client-server web-based editor using a thick client model
- The database is installed on a server and it is accessed online via the web
- The data is stored in a relational database
- Work tracking system available
- Ticketing system available

BulTreeBank CLaDA BG

System Specifics and Functionalities: the Utility

- 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
- The main form is literal-based (in other words, string-based)
- There is a possibility to consult a graphical representation, related information in the available dictionaries in the system

User Interface



BulTreeBank CLaDA BG

Supporting Language Resources

- In our view the necessary minimum of functionalities of such a system would include:
  - an editor of lexical entries supporting different structures of interrelated elements
  - access to existing dictionaries and corpora
  - concordances and other materials
- At the moment the integration of BTB-WN has been made with OEW, but a coverage comparison and linking between corresponding senses is planned to be done with several dictionaries such as
  - Bulgarian Explanatory Dictionary
  - Bulgarian Inflectional Lexicon (currently partially done)
- The system supports mapping to Wikipedia via the inclusion of Wikipedia article URI to the corresponding synset

Conclusions

- CLaDA-BG-Dict is an editor, which could be used both for creating lexical databases like wordnets as well as traditional types of dictionaries
- 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 13 000 synsets together with appropriate examples

Future Plans

- Our vision for future is to enhance replicability and re-usage of dictionary compilation as much as possible.
- In its current beta-version the system uses its own format for uploading corpora and other digitally-born or digitized dictionaries.
- However, it is planned to conform 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.

BulTreeBank CLaDA BG