Integrated Language and Knowledge Resources for a Bulgarian-Centric Knowledge Graph

Kiril Simov[0000-0003-3555-0179]

Linguistic Modelling and Knowledge Processing Department,
Institute of Information and Communication Technologies, Bulgarian Academy of Science,
Acad. G. Bonchev 25A, Sofia, Bulgaria
kivs@bultreebank.org

Abstract. This paper reports on the integration of language and knowledge resources within CLaDA-BG infrastructure. The idea is to encode linguistic knowledge on all levels of language starting from text, grammatical annotation and lemmatization to semantic and conceptual annotation. Our goal is to support conceptual annotation of various research objects (mainly texts). One of the main applications will be the management of a Bulgaria-centric Knowledge Graph.

Keywords: Language Resources, Language Technology, Knowledge Graph, Digital Humanities

1 Introduction

Here we present the first steps in the creation of language infrastructure for access and management of knowledge resources for cultural and historic heritage (CHH) resources. This infrastructure is part of CLaDA-BG.

CLaDA-BG is the Bulgarian National Interdisciplinary Research E-Infrastructure for Bulgarian Language and Cultural Heritage Resources and Technologies. In the spirit of European CLARIN and DARIAH, the mission of CLaDA-BG is to establish a national technological infrastructure of language resources and technologies (LRT), and cultural and historic heritage resources and technologies. The consortium of CLaDA-BG comprises 15 organizations including research institutes at the Bulgarian Academy of Sciences, several universities, the National Library "Ivan Vazov" in Plovdiv, and two museums. Thus the consortium includes not only technological partners, but also content and expertise providers many of which are simultaneously users of CLaDA-BG as a research infrastructure.

The main goal of the infrastructure is to provide public access to these resources and technologies for various societal tasks, targeted at wide audience. The infrastructure aims to support predominantly researchers in Art, Humanities and Social Sciences to process Bulgarian language texts and CHH datasets necessary for their research. The different types of objects of study, representation and search are integrated on the basis of common metadata categories and via textual descriptions. The language resources

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 9, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2019. ISSN: 1314-4006, eISSN: 2535-0366

and the textual descriptions of other objects are integrated with the help of a common Bulgaria-centred Knowledge Graph.

2 Research in Social Sciences and Humanities

In order to support the research within social sciences and humanities (SSH), CLaDA-BG needs to provide management of information of a huge variety of research objects including different kinds of texts (various genres, domains, time periods), artefacts models, art masterpieces representations and descriptions, etc. The top unification of these data is the metadata, but very little common information can be represented in this way. On the other end of the scale is the very specific data and tools for its management like creation (digitization), representation, generalization, search, etc. We consider as one of the step of doing research within SSH the identification of information of interest and its simultaneous observation within the same context.

In order to support this type of research in social sciences and humanities, we would like to put the varying types of data in the context of each other. This need was recognised during the first stage of the project. The approach for interlinking of the data was named *contextualization*.

The main characteristics of the contextualization are *time* and *space* – which events happened at the same time or in the same space. The additional characteristics include also the participants in the specific events; the similar constructions of physical objects like form, size, material; the similar style of representation in images, text and sounds; the same school of production; etc. Thus our motto is: *Everything in our world is connected and appears in a context*. Excluding the metadata which is obviously part of the context description most of the information within an individual dataset of CHH resources does not contain enough contextual information for integration with other such datasets. This is why we need to construct a new layer of information necessary to support contextualization.

In the course of implementing CLaDA-BG, we will develop the notion of contextualization in depth. As a starting point we focused on the following entities:

- **People** – their biographies – their characteristics, motivations, opinions, events in their lives, roles they played
- **Objects** – geographical, artefacts, etc. and their features
- **Events** – place, time, participants (People, Objects), relations to other events
- **Documents** – authors, content, opinions, mentions of people, events, entities, etc.

We consider text as the main source of information for the represented objects. As technology we consider Linked Open Data and the represented information will form a knowledge graph (Ehrlinger and Wöß 2016). Because the main body of knowledge represented within it will be related to Bulgaria, we call it *Bulgaria-centric Knowledge Graph* (*BGKG*).

Thus, the development of the different resources and technologies within CLaDA-BG will be guided via the construction of BGKG. One important element of knowledge within BGKG is the provenance statements about each piece of information. This will provide a basis for evaluation of the reliability of the represented knowledge.

The research within CLaDA-BG will be supported by the positioning of the searched information within the context of other types of information. This would help for the better understanding of the information of interest.

As it was mentioned above, an important element in the construction of BGKG is an appropriate set of language resources and technologies. We expect them to provide a mechanism for extraction of knowledge to be added to the graph from different textual resources. In the next section the currently available resources for Bulgarian and their integration are presented.

3 Language Resources and Technologies in CLaDA-BG

Within the CLaDA-BG plan for Language Resources and Technology we follow the notion of Basic Language and Resource Kit – BLaRK – (Krauwer 2003). Here is the initial list of the envisaged language resources. During the first phase of CLaDA-BG we focused on the consolidation and integration of existing language resources for Bulgarian, and partially for other languages. The resources will be published at the portal of CLaDA-BG in the requested format and with the respective metadata – post adaptation, since they were produced in previous projects. We are also working on the documentation and metadata descriptions. Here is a list of these resources:

- BulTreeBank in the original HPSG-based annotation (Simov et al. 2002)
- BulTreeBank in the Universal Dependency representation – several versions (Osenova and Simov 2015)
- BulTreeBank POS Corpus
- BulTreeBank WordNet (BTB-WN) – (Osenova and Simov 2018)
- Bulgarian CLEF Corpus – Question Answering and Information Retrieval. It comprises a corpus of about 20 million news items, a corpus of 1000 questions with marked answers, and 50 topics for Information Retrieval tasks.
- BulTreeBank Frequency List of word forms
- Bulgarian National Reference Corpus – BulTreeBank. The corpus consists of texts from the end of 20th and the beginning of 21st century – about 70 million running words. The corpus and the software will be partially updated during the first phase of CLaDA-BG
- BulTreeBank Stopword List

Additionally, we have worked on the creation of new resources as well as on the extension and integration of existing language resources. They are introduced below:

Large corpus of textual fragments

This corpus is under construction. More than 10 million documents are downloaded from the Bulgarian Internet. All n-grams from 1- to 6-grams will be extracted with their frequencies. Additionally, up to 2000 example contexts will be selected for each n-gram. Both data will be available for search and for download. The data will be useful for searching for patterns within contemporary Bulgarian texts and for the training of word embeddings.

Bulgarian Wordnet (BTB-WN)

BTB-WN is a WordNet of Bulgarian constructed along the lines of Princeton WordNet for English. The words are represented in synonymical sets (called synsets). Each synset represents one meaning per words in it. The meaning is described by a gloss and it is illustrated by example usages. It can be assumed that each synset represents a concept. The set of synsets is organized within a network via semantic and lexical relations like hyperonymy, meronymy, antonymy and others.

BTB-WN has been created via three different approaches: (1) by manual translation of English synsets from the Core WordNet subset of Princeton WordNet (PWN) into Bulgarian. This step ensures comparable coverage of the most frequent senses between the two WordNets; (2) by identification of senses used in the Bulgarian TreeBank (BTB). The identified senses have been organized in synsets for the BulTreeBank WordNet. The newly created Bulgarian synsets are mapped onto the conceptual structure of PWN. In this way, the BTB-WN was extended with real usages of the word meanings in texts. (3) by sense extension, which includes two activities: a) detection of the missing senses of processed lemmas in BulTreeBank and adding them to the BTB-WN, and b) a semi-automatic extraction of information from the Bulgarian Wiktionary, mapped to synsets from PWN and then manually checked.

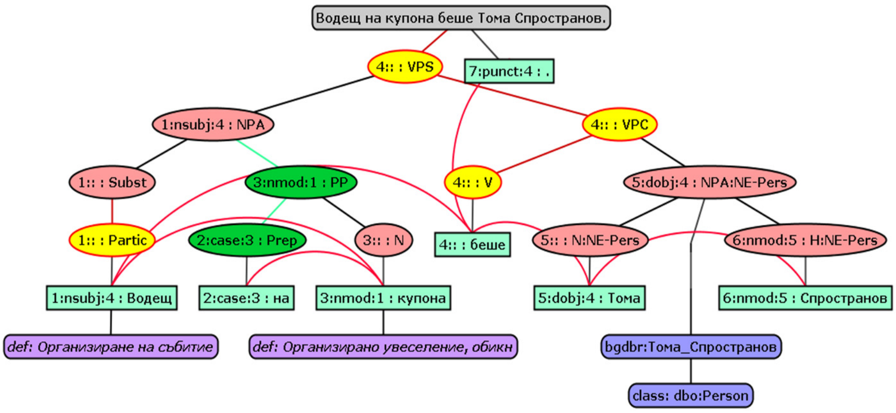
During the first phase of CLaDA-BG, the BTB-WN was manually extended by nearly 10 000 new synsets. The selection of the new words follows the BulTreeBank Frequency List. In this way, the coverage of BTB-WN was extended. Currently BTB-WN has more than 22 000 synsets.

Treebank of Bulgarian (BulTreeBank)

BulTreeBank is an HPSG-based treebank of Bulgarian. It contains a little more than 15000 sentences. The original annotation is constituent-based, with syntactic labels reflecting the type of the phrase (VPC = head complement verbal phrase). In each constituent the head is marked (if the phrase has a head). This marking of the head within phrases facilitates the conversion to a dependency format. The current version of the Treebank is represented in Universal Dependency Format and it is freely available from the UD website.[1]

During the creation of BTB-WN the treebank was annotated with senses from it (Popov et al. 2014). During the first phase of CLaDA-BG only some abbreviations were annotated with senses. In addition to common words, Named Entities (NEs) are annotated with four categories: people, locations, organizations, and others. We also have annotated them with URIs from DBpedia (Wikipedia) when the NE is in the Bulgarian Wikipedia. If not, they have been annotated with classes from the DBpedia ontology. Here is an example:

[1] https://universaldependencies.org/



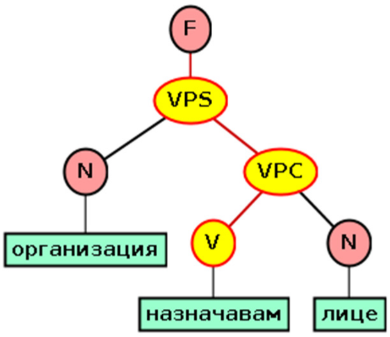
The sentence is: “*Водещ на купона беше Тома Спространов.*” (“The host of the party was Toma Sprostranov.”) The two open class words are connected with the respective synsets from BTB-WN, represented here by their definitions. The word “*Водещ*” (“The host“) is a participle of the verb “*водя*” (“to organize“) and it is annotated with that sense of the verb. From the fact that it is participle, present tense, it follows that the word denotes the person who is organizing the event. The word “*купона*” (“the party”) is connected to the definition “*Организирано увеселение …*” “(Organized entertainment”). The host was the disc jockey Toma Sprostranov who has a Wikipedia page:

`https://bg.wikipedia.org/wiki/Тома_Спространов`

In the image we have used the namespace `bgdbr:` defined at `http://prefix.cc/bgdbr`. In the cases when there is no Wikipedia page for the corresponding named entity we add only a class from the DBpedia ontology, such as `Person`, `Politician`, `Musician`, `Country`, `City`, `Document`, etc.

Valency lexicon

The Valency lexicon was constructed on the basis of annotations from BulTreeBank. After the extraction of verbs with their arguments from the treebank we classified the verbs by their senses within BTB-WN and then the arguments were also mapped to the corresponding synsets. Thus, one syntactic frame could result in several semantic frames. Here is one example:



It represents the case when an organization appoints a person. During the first phase we did not modify this lexicon, but with the available annotation of the treebank with senses from BTB-WN and thanks to the fact that the frames are extracted from the treebank, in the Treebank there is a connection of each verb with the corresponding frame. This allows the Treebank to be used for training of machine learning techniques to assign the correct frame in context for each verb.

Integration with Wikipedia, DBpedia, WikiData

Within CLaDA-BG, we plan to connect language resources with knowledge sources. In this way we will provide data for the implementation of language pipelines for annotation of textual data with world knowledge. Such an annotation will facilitate the extraction and integration of new knowledge within the BGKG.

The first step in this direction has been undertaken with the mapping of BTB-WN to the Bulgarian Wikipedia. We consider the mapping of two semantic objects – *concepts* (meanings expressed by common words) and *real-world things* (called named entities). The integration is meant to be verified manually in order to ensure the high quality of the result. The integrated knowledge graph will include the current version of BTB-WN extended with: a) new senses and new synonyms for existing synsets - all extracted from the articles in the Bulgarian Wikipedia; b) a controlled number of named entities that are specific to Bulgaria and c) the number of terminological concepts in various domains. Thus the integrated resource will combine general lexica with encyclopaedic knowledge (terminology). We have selected over 13000 Wikipedia pages that share the same lexical units with BTB-WN. Currently about 2000 of them are already processed. We envisage to complete the task by September 2019 – see (Simov et al. 2019).

The integration of language resources and Wikipedia will be used at least in two directions: (1) training of a wider set of processing modules, performing multiple tasks simultaneously in an end-to-end training fashion, and (2) contextualization, through the relations from the text with the encyclopaedic information. The latter is considered very important for the connection between language processing and suitable information extraction from textual descriptions of cultural and historical objects.

Wikipedia plays an important role in the construction of many knowledge graphs such as DBPedia[2] and Wikidata.[3] Thus, mapping from BulTreeBank and from BTB-WN to Wikipedia provides direct access to them. We consider such a mapping between a knowledge graph and language resources obligatory in order to support the creation, the access, and the usage of the knowledge graph. In the next section we outline some usages of the integrated language resources in relation to knowledge graphs.

4 Language Technologies for Knowledge Graphs

We aim at creating a semantically integrated environment for maintaining possibilities of referring to texts and descriptions of cultural or historic objects. For this purpose the

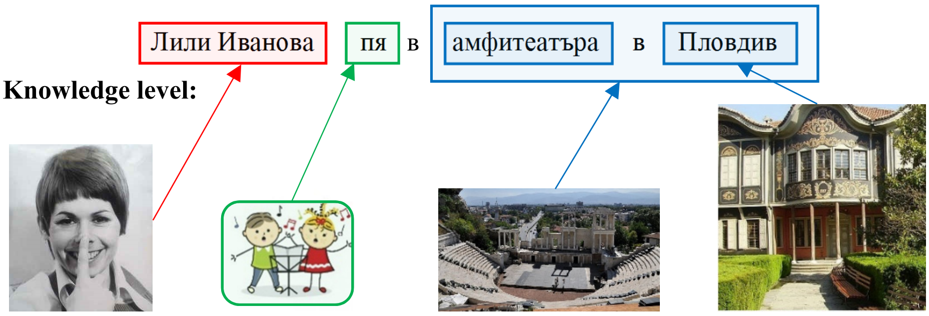
[2] https://wiki.dbpedia.org/

[3] https://www.wikidata.org/wiki/Wikidata:Main_Page

texts and descriptions of collections will be annotated with an integrated ontology (including DBpedia ontology, but also Lemon ontology, domain ontologies, provenance ontology, etc.) and then the annotation will be uploaded into an RDF repository.

The main application of the integrated language and knowledge sources is the semantic annotation of texts with conceptual information from the knowledge graph. For example, if we have the sentence "Лили Иванова пя в амфитеатъра в Пловдив" ("Lili Ivanova sang in the amphitheater in Plovdiv") we expect the language technologies to be able to recognize the involved named entities and the event described by the sentence. The following picture depicts the desirable analysis.

Language level:



On the language level the language technology tool recognises a person name "Лили Иванова" ("Lili Ivanova"), the verb form "пя" ("sang"), the noun form "амфитеатъра" ("amphitheater") and the location name "Пловдив" ("Plovdiv"). Besides these, the tool will provide the syntactic structure of the sentence.

The integration of the language resources and the knowledge graph provides the following information: the person Lili Ivanova is the famous Bulgarian singer represented in the knowledge graph; the location Plovdiv is related to the description of the town of Plovdiv; in Plovdiv there are two amphitheaters, but only one of them is used for concerts, thus the phrase "амфитеатъра в Пловдив" ("the amphitheater in Plovdiv") will be connected to the correct one; the event denoted by the whole sentence is determined by the verb in the appropriate sense. Such a semantic annotation service has many applications:

- **Extraction of new knowledge from text**. If the above sentence is in a news article with appropriate metadata we could add facts like: Lili Ivanova was in Plovdiv on certain date; she had a concert in Plovdiv; the concert was in the amphitheater; etc.
- **Querying the knowledge graph**. The users will be able to formulate their queries to the knowledge graph in the form of connected text. If we ask when Lili Ivanova sang at the amphitheater in Plovdiv, then the above analysis is easy to convert to the appropriate formal query for the knowledge graph.
- **Indexing of text corpora**. The segments of text linked to the entities or facts in the knowledge graphs can be easily indexed by them.

We think that the applications are much more and in future work we will explore them. In addition to them the possible reasoning over the knowledge graph will be explored like geographical one: *Lili Ivanova sang in the amphitheater in Plovdiv* implies that *she sang in Plovdiv*, *in Bulgaria*, and so on.

5 Conclusions

In this paper we present the current state of construction of services based on integrated language and knowledge resources. We consider them as a basis for implementation of language technology services to support the link between the various datasets necessary for supporting SSH research and the common knowledge necessary to integrate them into a knowledge graph.

Within CLaDA-BG infrastructure are working towards construction of Bulgarian-centric Knowledge Graph based to available knowledge graphs on the web, but extended with facts related to Bulgaria extracted from the data provided by the partners within the consortium.

Acknowledgements

This research was partially funded 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 DO01-164/28.08.2018.

References

Ehrlinger, L., & Wöß, W. (2016). Towards a Definition of Knowledge Graphs. *SEMANTICS 2016: Posters and Demos Track*. September 13-14, 2016, Leipzig, Germany

Krauwer, S. (2003). The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of 2nd International Conference on Speech and Computer (SPECOM2003).*

Osenova, P. & Simov, K. (2015). Universalizing BulTreeBank: a Linguistic Tale about Glocalization. In *The 5th Workshop on Balto-Slavic Natural Language Processing. Hissar*, Bulgaria. pp. 81-89.

Osenova, P. and Simov, K. (2018). The Data-driven Bulgarian WordNet: BTBWN. *Cognitive Studies | Études cognitives*. vol. 18(1713).

Popov, A., Kancheva, S., Manova, S., Radev, I., Simov, K., Osenova, P. (2014). The Sense Annotation of BulTreeBank. *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, 2014.

Simov, K., Osenova, P., Slavcheva, M., Kolkovska, S., Balabanova, E., Doikoff, D., Ivanova, K., and Simov, A., Kouylekov, M. (2002). Building a Linguistically

Interpreted Corpus of Bulgarian: the BulTreeBank. *LREC 2002*. Las Palmas, Canary Islands - Spain. pp. 1729-1736.

Simov, K., Osenova, P., Laskova, L., Radev, I.,, and Kancheva, Z. (2019). Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia. *The 10th Global WordNet Conference.* Wroclaw, Poland.

Received: June 15, 2019
Reviewed: July 05, 2019
Finally Accepted: July 20, 2019