Integrated Language and Knowledge Resources for CLaDA-BG

Kiril Simov
LMaRK
IICT-BAS, Bulgaria
kivs@bultreebank.org

Petya Osenova
LMaRK
IICT-BAS, Bulgaria
petya@bultreebank.org

Abstract

This paper presents the envisaged integration of the language resources for Bulgarian with the knowledge sources like ontologies and linked open data to support their joint usage with respect to the cultural and historical heritage (CHH) objects. We started with the knowledge integration of the language resources for Bulgarian. Our plan is to continue with the addition of selected CHH objects to the initial integrated data. Based on the available Bulgarian resources like dictionaries and corpora as well as on the Bulgarian Wikipedia, DBpedia and Wikidata, we have constructed the first version of a Bulgaria-centered Knowledge Graph. It represents the conceptual information for the Bulgarian virtual infrastructure CLaDA-BG.

1 Introduction

Nowadays vast networks with linked objects are dominant in many areas of life, including tools and data in NLP. Among the many prominent initiatives in linking available data in various combinations are the following: CLARIN-ERIC[1] (the infrastructure that combines the strengths of the language resources and technologies), the Linked Open Data Cloud[2] (language resources with ontologies), the Predicate Matrix[3] (a lexical resource that integrates the information from different semantic and syntactic resources such as FrameNet, VerbNet, PropBank, WordNet), BabelNet[4] (a knowledge base with a strong multilingual value), PARTHENOS project[5] (integrates cloud storage with services and tools and support collaborative working on language and CHH data), SSHOC[6] (connecting existing and new infrastructures from the SSH ERICs), ELEXIS[7] (the European Lexicographic Infrastructure) and many others. All these projects and initiatives focus on the idea of linking. This means: linking tools, resources, architectures, knowledge bases and infrastructures. Our work has many lines in common with the mentioned projects. Similarly to CLARIN-ERIC and PARTHENOS, we aim at providing adequate language technology for Linguistic Studies, Humanities, Cultural Heritage, History and related fields, making them mutually understandable and coherent. As in Predicate Matrix, BabelNet and ELEXIS, we combine information from various resources - in our case these are BTB-WordNet, Valency dictionary, Wikipedia, Wiktionary. The differences can be summarized as follows: our focus is set particularly on Bulgaria-related data; apart from integrating resources, we rely on annotated biographical data from historians, librarians, museum workers; our semantic and encyclopaedic resources for Bulgarian do not have the coverage of English or German related ones. Thus, we rely on getting more knowledge through the cross-lingual mappings coming from wordnets, Wikipedia, etc. In the future we will investigate the

[1] https://www.clarin.eu/

[2] https://lod-cloud.net/

[3] http://adimen.si.ehu.es/web/PredicateMatrix

[4] https://babelnet.org/

[5] http://www.parthenos-project.eu/

[6] https://sshopencloud.eu/

[7] https://elex.is/

Kiril Simov and Petya Osenova 2020. Integrated Language and Knowledge Resources for CLaDA-BG. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 137–144.

possibility to incorporate in our work also ideas from VerbAtlas - di Fabio et al. (2019). VerbAtlas is a manually-crafted verbal semantic resource structured into frames. It groups semantically-coherent synsets from WordNet. It is the first resource enriched with semantic information about implicit, shadow and default arguments following Pustejovsky (1995). VerbAtlas aims at improving the main features of the existing verbal inventories (FrameNet, PropBank, VerbNet), while also adding new semantic information.

CLaDA-BG[8] is the Bulgarian National Interdisciplinary Research E-Infrastructure for Bulgarian Language and Cultural Heritage Resources and Technologies. In contrast to other EU infrastructures that started separately as CLARIN and DARIAH, and later on in some countries (Austria, the Netherlands, Greece and others) combined or started to work in a closer cooperation, in Bulgaria the joint infrastructure started as a joint endeavour from the beginning. 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), as well as cultural and historical heritage (CHH) resources and technologies in a connected framework. 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 does not include only technological partners, but also content providers and experts in history, library studies, arts.

The main goal of the infrastructure is to provide public access and an integrated version of the available resources and technologies for various societal tasks, targeted at a wider audience. The infrastructure aims to support primarily researchers in Art, Humanities and Social Sciences to process Bulgarian language texts and CHH datasets necessary for their research. However, the real applications are envisaged to go beyond the research framework, since many areas can profit from linked knowledge. Thus, the results will be applicable also to education and industry.

Needless to say, linking data in a broader sense is a challenge. First of all, due to the fact that data itself is diverse. Second, because these data exist in various formats and representations. Third, different tools are necessary for manipulating the data. Last but not least, the data has to be made to communicate to each other, since it supplies similar or different pieces of knowledge that might contradict or remain incomplete thus leading to misunderstanding or wrong assumptions.

For all the reasons mentioned above, we focus on putting the varying types of data into the context of each other. The approach for interlinking of the data is called *contextualization*. 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 and the textual descriptions of other objects are integrated with the help of a common Bulgaria-centred knowledge graph - *BGKG*. Thus, the language description has become the main brick for creating the knowledge graph. We also plan to integrate links to images and digitized/3D-scanned objects.

The existing open data for Bulgarian, such as Wikipedia, DBpedia and Wikidata[9] are still scarce and/or not completely reliable. For that reason, we provide a) linking with our in-house lexicons and corpora, and b) gather data from our content providing partners that are of high quality.

In this paper we present the core sets of language resources that in our view are necessary to support research in social sciences and humanities. We also show how they are integrated in order to support the semantic annotation of texts with conceptual information from the knowledge graph with the aim to ensure: extraction of new knowledge from text, querying over the knowledge graph, and indexing of texts within the CLaDA-BG repository.

2 Integrated Bulgarian Language Resources

Since it was decided that language data and language descriptions of library/museum objects will be the connecting parts within the CLaDA-BG, we have to ensure the necessary framework and technology. For the necessary language resources set as a prerequisite we rely on the Basic Language and Resource Kit – BLaRK – Krauwer (2003). Below is the initial list of these language resources. It has to be noted that this data has been available for Bulgarian for years, but they have to reach the designated size in the

[8] http://clada-bg.eu/

[9] Here we consider the part of Wikidata that is equipped with Bulgarian labels.

brackets and to become easily searchable on the web and within the CLARIN repository. The basic language resources are:[10]

Corpora

- Text Archive for Bulgarian (minimum 100 million running words);
- Morphologically Annotated Corpus (1 million running words);
- Syntactically annotated corpus (1 million running words);
- Semantically annotated corpus with ontological and fact information (1 million running words);
- Domain corpora (minimum 100 000 running words per domain)

Lexicons

- Bulgarian Wordnet (BTB-WN) (50 000 synsets of coverage of the lemma senses in a related semantically annotated corpus)
- Valency lexicon (coverage of the verbs in BTB-WN)
- Domain dictionaries (minimum 100 000 running words per domain)
- Representative lists of Bulgarian names (coverage of the names of the public figures, location and organization names. Additionally, they will include relations to the Bulgarian Wikipedia)

The language processing tools include minimally the following ones: morphological, shallow syntactic, deep syntactic, and semantic analyzers, named entities recognition and identification modules.

During the first year of the CLaDA-BG project we focused on the integration of the various existing language resources and performed only minimal extension in order to make them usable.

As a basis for the manually annotated corpus we use the texts included in BulTreeBank - an HPSG-based treebank for Bulgarian - comprising about 260 000 running tokens. These texts were annotated before the start of CLaDA-BG with senses from BTB-WN and instances from DBpedia, URLs from WikiPedia and classes from DBpedia ontology (see Popov et. al, (2014)). The original annotation is an HPSG-based constituent structure with marked up the head in each phrase. It was converted automatically to a dependency format. The dependency annotation follows the Universal Dependency guidelines.[11] The original Treebank is also manually annotated on morphosyntactic level with a rich tagset (680 tags – Simov et al. (2004)) and lemmas.

The BTB-WN currently contains 22 000 synsets which cover all the words within BulTreeBank and most frequent words over the Bulgarian national reference corpus (about 100 million running words). We started with the extension of the information within BTB-WN by adding inflectional paradigms to each lemma in the synsets and with their mapping to articles from the Bulgarian Wikipedia – see Simov et al. (2019). The inflectional information is important because many lemmas in BTB-WN can belong to different inflectional paradigms. The information from Bulgarian Wikipedia provides not only additional encyclopedic information for the named entities, but also a terminological one. During the process of mapping BTB-WN to Wikipedia, new senses have been added to the lexical resource. In addition, the mapping to Wikipedia provides a source for new relations between the synsets in BTB-WN. On the basis of the current synsets in BTB-WN, we extracted about 13 000 Wikipedia articles. These articles were manually inspected and mappings between the synsets and the articles were established. The mapping follows the approach of McCrae (2018). In addition to the Wikipedia articles that correspond to the synsets in BTB-WN, we selected and extracted 10 899 Wikipedia articles that relate to the names in a Bulgarian gazetteer. This gazetteer represents the most important names in the Bulgarian National Reference Corpus. From them 1 515 pages were already extracted on the basis of the lemmas within BTB-WN. The remaining 9 384 pages were classified as Bulgarian locations, other locations, people, organizations, and other. In this way we extend BTB-WN with important information for Bulgaria named entities.

[10] In brackets we put the desirable minimal size of the corresponding resource that would make it applicable in many areas of usage.

[11] https://universaldependencies.org/

The creation of the Valency lexicon started by generalization over the annotations within 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 on Fig. 1:

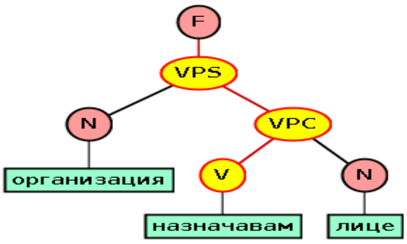

Figure 1. An example of a generalized syntactic frame.

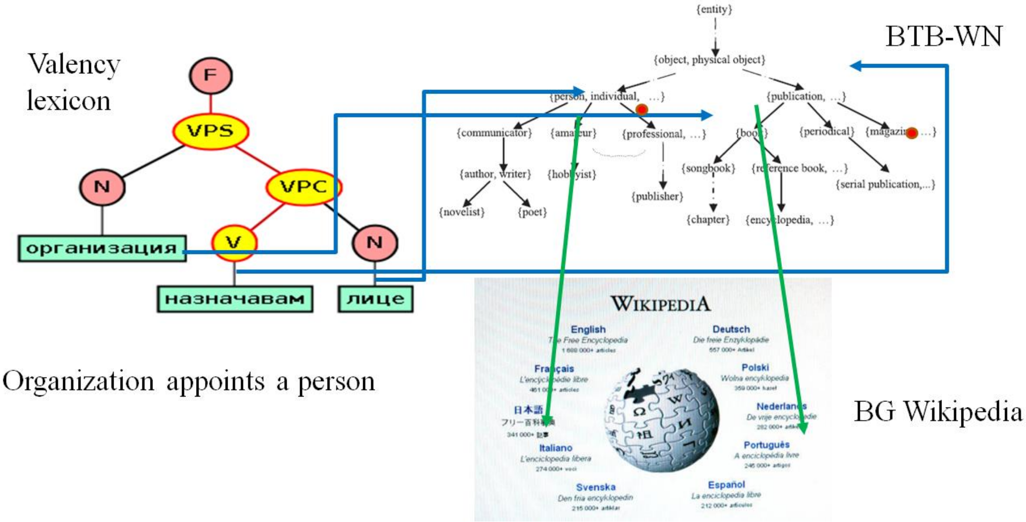

Figure 2. Mapping between the Valency lexicon, BTB-WN and BG Wikipedia.

It represents the case when an *organization* (subject noun - N) *appoints* (V) *a person* (object noun - N). During the first phase we did not modify this lexicon, but focused on ensuring the connection of each verb with the corresponding frame. This was possible since the treebank was annotated with senses from BTB-WN and the frames were originally extracted from the treebank. Later on, a number of missing senses were added as well. The rich annotation of the treebank allows it to be used for the training of machine learning techniques that assign the correct frame for each verb depending on the context. In the next phases of CLaDA-BG we plan to extend the Valency lexicon to also cover the verbs within BTB-WN. On the other hand, the mapping between BTB-WN and encyclopedic knowledge ensures a mapping between the Valency lexicon and encyclopedic knowledge. Currently, the encyclopedic knowledge is being extracted from Wikipedia, DBpedia, Wiktionary as well as expertise data (biographies of important Bulgarian people, descriptions of significant events, etc.) but during the next phases of the project it is envisaged to cover the whole knowledge graph. Thus, on the lexical level we have the mappings between the Valency lexicon, the Bulgarian Wordnet (BTB-WN) and Bulgarian Wikipedia as depicted on Fig. 2 above.

On the corpora level all the information needed for an end-to-end representation was integrated: tokens, grammatical annotation, lemmatization, syntax, word senses and Named Entity categories. Fig. 3 shows an example of a sentence annotated with senses from the Bulgarian Wordnet BTB-WN and URLs from Bulgarian Wikipedia on top of the morphosyntactic analysis.

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 is annotated with that sense of the related verb. From the fact that it is a 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 DJ Toma Sprostranov who has a Wikipedia page:

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

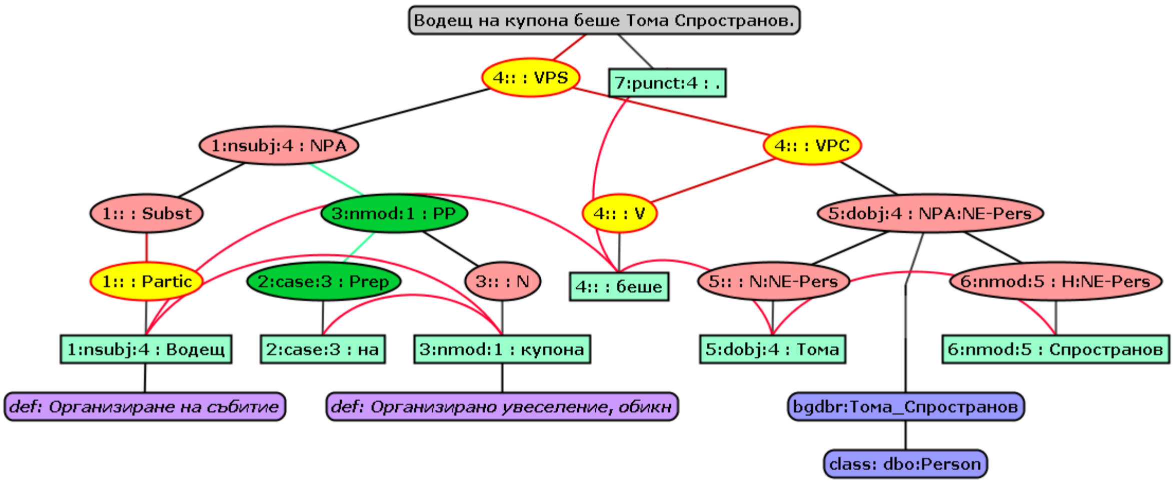

Figure 3. An example of a sentence annotated with morphosyntactic and semantic annotation.

In the image we 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. These annotations provide access from the treebank to the knowledge within BTB-WN and Wikipedia (later from the knowledge graph). Thus, the users of these integrated resources have access to the knowledge not only in the annotated corpus, but also within the lexicons and encyclopedic resources. The annotation of all these language levels over the same text documents provides a good basis for the widely used end-to-end neural models.

The integration of the language resources will be used at least in two directions (1) training of a wider set of processing modules, and (2) contextualization through the relations from the text to 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.

The integration of language resources and encyclopaedic knowledge is the first step in the direction of constructing a knowledge graph for CLaDA-BG aligned to language resources for Bulgarian.

3 Towards a Bulgaria-Centric Knowledge Graph

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 texts and descriptions of collections should be first annotated with an appropriate ontology, and then the annotation should be uploaded into an RDF repository. The first version of BGKG is based on the Bulgarian DBpedia knowledge graph. In the process of implementation of CLaDA-BG we will gradually add knowledge from other sources. Besides Wikipedia and DBpedia we envisage the inclusion of Wikidata as part of the initial knowledge graph. The integration of these sources of knowledge is guaranteed by their design. Wikidata as a knowledge source is considered with a higher level of quality because it follows rigorous rules for the construction and manual inspection phases.

As one step in the process of doing research within social sciences and humanities (SSH) we consider the identification of information of interest and its simultaneous observation within the same context. In order to support the research within 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), artefact models, art masterpieces representations and descriptions, etc. The top unification of this data is the metadata,of course, but in fact very little common information can be represented in this way. In order to escape from the problem, we consider a new layer of information between the metadata and the actual datasets within SSH. Fig. 4 depicts the architecture we want to achieve.

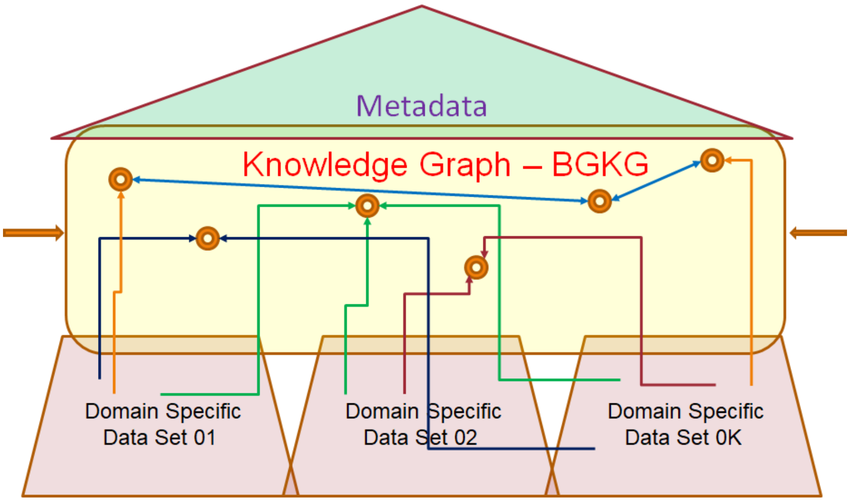
Metadata
Knowledge Graph – BGKG
Domain Specific Data Set 01
Domain Specific Data Set 02
Domain Specific Data Set 0K

Figure 4. The knowledge graph abstract architecture.

The newly introduced layer comes between the metadata layer where information about the objects is represented within the domain datasets. In many cases the researchers that use metadata to find the necessary information have to know additional information about the content of the domain data. Also, they need to find accordingly this information in the domain datasets. In order to provide such a functionality, we propose to construct the *Bulgarian-Centric Knowledge Graph* (BGKG) for supporting access to heterogeneous datasets. The approach for interlinking of these data was named *contextualization*. The aim of the integrated language resources is to provide a language layer for accessing the BGKG.

The main characteristics of the contextualization are time and space – what events happened at the same time or in the same space. The additional characteristics include also: the participants in a given event; 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*. The knowledge graph is a network of interlinked descriptions of people, events, geographical entities, objects, documents, authors, opinions, etc.:

- People – biographical data – events in their life, their roles
- Geographical entities – history of cities, etc.
- Objects – creation, materials, form, discovery
- Events – place, time, participants, connection to other events
- Documents – authors, contents, opinion about peoples, events, …

As already mentioned above, we consider the text as the main source of information for the represented objects. Linked Open Data and knowledge graphs were chosen for its representation Ehrlinger and Wöß (2016).

We first started with the existing DBpedia knowledge graph, constructed on the basis of Bulgarian Wikipedia. This knowledge graph is used currently in two ways: (1) for the semantic annotation of a huge web-based corpus; and (2) as an initial source of identifiers and facts for the construction of BGKG.

For the application in (1) we rely on the existing NLP pipeline for Bulgarian to annotate the documents within the web-based corpus mentioned above with Named Entities and identifiers from the knowledge graph. This will allow searching via entities from the knowledge graph.

The usage (2) of the initial knowledge graph is to support the creation of BGKG. The approach we selected relies on knowledge extraction from text. Thus, we started the creation of a semantically integrated environment for maintaining possibilities of referring to texts and descriptions of cultural or historic objects. For this purpose, the texts and descriptions of cultural or historical collections should be annotated with an appropriate ontology, entity identifiers and then the annotation should be converted to RDF triples and uploaded into an RDF repository. The whole process includes the following steps:

- Selection of appropriate ontologies
- Mapping of the ontologies to the integrated language resources
- Semi-automatic annotation of domain documents with Named Entities, their identifiers, concepts, relations between them
- Extraction of RDF triples from the annotations
- Manual assessment of the extracted facts and adding them to BGKG

The selection of the ontologies to be used in the creation of BGKG depends on the data available to the partners within CLaDA-BG. Each selected ontology will be aligned to the Bulgarian Wordnet BTB-WN. This will provide a better understanding of the ontology through appropriate lexicalizations. The classes of the ontology will be aligned to the synsets within BTB-WN. The properties will be aligned to triples of synsets (one or more triples of this kind) where the properties correspond to event synsets (usually verbal, but noun and adjectival synsets are also possible) and the subject and object of the properties will be mapped to the relevant synsets. In this way, the related properties will be mapped to the same event. For example, properties like `date-of-birth`, `place-of-birth`, `mother-of`, and `father-of` will be mapped to the synset for the event `birth`. The alignment of the ontology to BTB-WN will give the possibility of automatic processing by the available NLP pipelines. The actual integration could also require additions to the inflectional lexicons for the new words as well as annotation of new texts.

The annotation of new texts will be done semi-automatically, thus including human inspection. Human intervention will undergo various changes during the annotation process. At the beginning it will be during the entities annotation, the identification selection, and the relation annotation. When there are enough manually annotated documents and a subsequent improvement of the automatic annotation, the human attention will be directed to the extracted facts for the knowledge graph.

The knowledge graph will be available via a search tool. The search tool will provide the following search possibilities: a) concept search; b) facet search (integrating several concepts) and c) combined search (integrating concepts with random key words). This will ensure similar search for mentions of conceptual information in the tests and in the semantic description of the cultural and historical objects. The inclusion of the language, cultural and historical information into a common knowledge graph will provide one of the main mechanisms to support the research through the contextualization of each object of analysis.

4 Conclusion

In the paper we present the ongoing development of language and semantic resources within CLaDA-BG to support research in Humanities and Social Studies. This is done through the exploration of text corpora and the description of cultural and historical objects. In the area of language resources, the integration of Bulgarian language resources through BTB-WN and the Bulgarian Wikipedia provides a basis for training of text indexing with instances and classes from a Knowledge Graph. The actual knowledge graph is based on DBpedia and Wikidata (including also information from Wikipedia). Besides the textual information, descriptions of cultural and historical heritage objects are expected to be mapped to the knowledge graph as well. This step will allow a joint search including a SPARQL endpoint. When developed enough, the knowledge graph will be provided freely for download as a linked open dataset.

The initial knowledge graph will be further extended by specific ontologies for modelling the specific classification schemata, or time and space, events, facts.

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 DO01-272/16.12.2019.

References

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

Andrea Di Fabio, Simone Conia and Roberto Navigli. 2019. *VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling.* Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), Hong Kong, China, November 3-7, 2019, 627–637.

Steven Krauwer. 2003. *The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap*. In Proceedings of the 2nd International Conference on Speech and Computer (SPECOM2003), 8–15.

John P. McCrae. 2018. *Mapping WordNet Instances to Wikipedia*. Proceedings of the 9th Global WordNet Conference (GWC 2018), 62–69.

Alexander Popov, Stanislava Kancheva, Svetlomira Manova, Ivajlo Radev, Kiril Simov and Petya Osenova, 2014. *The Sense Annotation of BulTreeBank*. Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13), 2014, 127–136.

James Pustejovsky. 1995. The Generative Lexicon. MIT Press, Cambridge MA.

Kiril Simov, Petya Osenova, and Milena Slavcheva. 2004. *BTB-TR03: BulTreeBank Morphosyntactic Tagset*. BulTreeBank Project Technical Report № 03. 2004.

Kiril Simov, Petya Osenova, Laska Laskova, Ivajlo Radev, and Zara Kancheva. 2019. *Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia*. Proceedings of the 10th Global WordNet Conference. Wroclaw, Poland, 290–297.