Chapter 7
Language Report Bulgarian

Svetla Koeva

Abstract This chapter reports on the current status of technology support for Bulgarian and highlights certain gaps. The analysis is based on the services and resources available in the European Language Grid in early 2022. While the LT field as a whole has significantly progressed in the last ten years, we conclude that there is still a yawning technological gap between English and Bulgarian, and even between German, French, Italian, Spanish and Bulgarian. It is exactly this distance that needs to be ideally eliminated, if not at least reduced, in order to move towards Digital Language Equality for Bulgarian.

1 The Bulgarian Language

Bulgarian is the official language of the Republic of Bulgaria. It is spoken by over eight million native speakers. According to an assessment by the National Statistical Institute for the 2021 census, the population of Bulgaria is about 6,500,000. A report by the World Bank states that about 1.7 million Bulgarians lived abroad in 2020.

The official alphabet is Cyrillic. Bulgarian was the first Slavic language to have its own writing system, which dates from the 9th century. Bulgarian belongs to the family of South Slavic languages and forms part of the Balkan linguistic union. Bulgarian exhibits a number of specific characteristics that contribute to the richness of the language but can also be a challenge for natural language processing (NLP), e. g., a rather flexible word order, combined with the lack of morphological distinction for nominal cases and regular subject omission.

The Bulgarian constitution states that Bulgarian is the official language in the Republic of Bulgaria. All education and teaching provided as part of the current state curriculum, from preschool to university, is in Bulgarian. The Institute for Bulgarian Language of the Bulgarian Academy of Sciences is the official institution that monitors changes in the Bulgarian language, determines literary norms and reflects these changes in both orthography and grammar.

Svetla Koeva
Institute for Bulgarian Language Prof. Lyubomir Andreychin, BAS, Bulgaria, svetla@dcl.bas.bg

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_7

According to W3Techs, Bulgarian accounts for just 0.1% of the language content on the web (as of November 2021). Bulgarian internet users in 2020 increased by 31% in comparison to 2007 and already 46% of the total population uses the internet. Bulgarian Wikipedia, as an important source of data for NLP, has a considerably smaller size than the biggest Wikipedias.

Bulgaria's membership in the EU, together with the ideas of unity and diversity, and globalisation while preserving national identity, provides a real opportunity for the equal use of Bulgarian together with the other major European languages.

2 Technologies and Resources for Bulgarian

Language Technology (LT) provides solutions for the following main application areas: Text Analysis; Speech Processing; Machine Translation; Information Extraction and Information Retrieval; Natural Language Generation; and Human-Computer Interaction. This study on LT for Bulgarian is based mainly on the European Language Grid as of February 2022 (Koeva and Stefanova 2022).

Technological developments in recent years have enabled the processing of huge amounts of language data, and allowed the application of complex models and algorithms, which will lead to significant progress (including for Bulgarian). Bulgarian is present in several monolingual and multilingual corpora. Some of the multilingual corpora are sentence-aligned, which allows for cross-lingual research. However, large multilingual corpora are usually created automatically from the internet (often from Wikipedia). Annotated corpora with manually validated or manually assigned linguistic information are smaller in number and volume. There are very few examples of multimodal corpora. Among the multilingual annotated corpora where Bulgarian is present, there are two relatively large collections: Universal Dependencies treebank v2.8.1, and the annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1), both freely available. There is an expanding collection of datasets and models for Bulgarian (at Hugging Face).

Bulgarian is relatively well-resourced when it comes to dictionaries and thesauri. Most dictionaries have been developed at the Institute for Bulgarian Language, but due to copyright restrictions, some of them only offer single user queries or access for research purposes only. Parts of the Bulgarian WordNet are available for download, extended with semantic classes, new semantic relations and semantic frames.

There are several NLP libraries providing sets of linguistic annotations for Bulgarian (tokenisation, sentence splitting, paragraph detection, lemmatisation, named entity recognition, dependency parsing, etc.). A number of libraries provide deep learning techniques and knowledge graphs, and report good levels of accuracy and speed (e. g., Spark NLP). Recently, two NLP pipelines (including a tokeniser, a sentence splitter, a tagger, a lemmatiser and a dependency parser) have become available: UDpipe and NLP-Cube, trained for languages with UD Treebanks, including Bulgarian.

Generally, LTs for Bulgarian still dominate text analysis while multimodal input data (such as simultaneous text, images, audio and video) is rarely processed.

The quality of speech technology for Bulgarian is not yet satisfactory. There are still no accessible and reliable speech-to-text systems for Bulgarian, especially working in real time. Excluding the automatic translation offered by multinational enterprises, there are other available MT systems from and into Bulgarian with different types of access. The assessment of the quality of existing MT services, the number of language pairs, and the coverage of thematic domains still determines MT technologies for Bulgarian as underdeveloped. Recently, there have been serious advances in research on information extraction for Bulgarian: event extraction, sentiment analysis, fake news detection, fact-checking.

There is no dedicated funding or infrastructure for Bulgarian LTs. Many of the achievements and advancements in the development of language data and tools for Bulgarian have been the result of short-term funded projects and PhD theses.

A number of Bulgarian LT companies are very successful, for example, Ontotext, operating in the field of semantic technologies with its product GraphDB.

When we compare the two studies – Blagoeva et al. (2012) and Koeva and Stefanova (2022) – we can see that there is a development in LRs and LTs for Bulgarian, but this is also true for other European languages. Furthermore, in a comparative analysis in 2012, Bulgarian was ranked 15th in terms of technological support, while it is ranked 21st in 2022. Nowadays, technological progress is rapid, and we should consider language models such as GPT-3 and its successors for Bulgarian and the other European languages, which will necessitate significant investments.

3 Recommendations and Next Steps

Many commonly used AI technologies are still not available for Bulgarian (Human-Computer Interaction, multimodal processing, etc.), while for others, if technology has made advances, there are no available applications (summarisation, question answering, etc.). Progress is typically made abroad and Bulgarian is part of some multilingual systems for MT and speech analysis. There is already a need for open real time MT services from and to Bulgarian combining text and speech, taking into account context, communicative purposes and different environments. Thus, speech and text technologies for Bulgarian have to be combined with technologies for other modalities: real time image and video processing working simultaneously in multilingual environments. Natural language understanding and generation of Bulgarian have to become part of multilingual and multimodal processing.

Digital Bulgarian needs large-scale, long-term support, harmonised with the support for all European languages. The sporadic funding of various tasks and particular languages should be replaced by common goals and objectives for all European languages, which if provided with the necessary funding will lead to vast improvements. Efforts cannot be focused only on Bulgarian or on any single language, because multilingual and multimodal resources and technologies are currently needed.

A BLARK-like (Basic Language Resources Kit) minimum set of LRs and LTs for all European languages should be developed and maintained, taking into account that the minimum requirements change rapidly. In 2022, this set should contain large integrated models for as many applications as possible: real-time, multimodal, cross-domain and multilingual LRs and LTs; and a variety of domain-specific datasets.

Convenient and well-regulated access to data is essential for the development of new products, applications and services. To achieve a significant advance over the current situation, an increase of available (open and copyright-free) data for Bulgarian and other European languages is needed, as is an improvement in the legal conditions for (re)using data at the European level.

There is a need for dedicated education and training programmes in the field of LT and AI, as it has proven difficult to source researchers, linguists or engineers with the right combination of skills (e. g., Bulgarian language, computer science, linguistics).

To avoid the reduplication of efforts and to promote data-sharing, it is needed to strengthen and reinforce the European hubs and repositories, such as ELG, intended for ready-to-use datasets, models, tools and services. This will increase the overall language support and ensure the sustainability of LT solutions.

To conclude, although a number of technologies and resources for Bulgarian exist, there are far fewer technologies and resources for Bulgarian than for English as well as for some other European languages. Our vision is high-quality LT for all European languages that supports political and economic unity through cultural diversity.

References

Blagoeva, Diana, Svetla Koeva, and Vladko Murdarov (2012). *Българският език в дигиталната епоха – The Bulgarian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/bulgarian.

Koeva, Svetla and Valentina Stefanova (2022). *Deliverable D1.5 Report on the Bulgarian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-bulgarian.pdf.