Chapter 6
Language Report Bosnian

Tarik Ćušić

Abstract It is objective to state that there are no language technologies for the Bosnian language or initiatives for the digitalisation of the Bosnian language. Therefore, it is necessary to take initial steps towards technological support for the Bosnian language, in order to prevent its digital extinction. In Bosnia and Herzegovina, no programmes aimed at the research and development of language technology products have been initiated. The Bosnian language is present in the digital sphere more or less as much as it is included in foreign, multilingual tools and resources, which are mostly related to Machine Translation (Google Translate and others).

1 The Bosnian Language

The Bosnian language belongs to the West-South Slavic subgroup of the Slavic branch of the great Indo-European linguistic family. Bosnian has about 2.5 million native speakers in Europe. It is the official language in Bosnia and Herzegovina, along with Croatian and Serbian, where it is spoken by 1.87 million people, or 53% of the population. Bosnian is the native language of Bosniaks in Bosnia and Herzegovina, but also of members of other ethnic groups. Outside of Bosnia and Herzegovina, Bosnian is one of the official languages in Montenegro. Bosnian is also an officially recognised minority language in Croatia, Serbia, North Macedonia and Kosovo. In Western Europe and North America, Bosnian is used by about 150,000 people, and by 100,000 to 200,000 people in Turkey.

There is no single language law in Bosnia and Herzegovina that regulates the issue of official language use. However, Bosnian (along with Croatian and Serbian) is listed as one of the official languages in laws and regulations on primary education, secondary education and higher education.

Two writing systems are used in the Bosnian language: Latin and Cyrillic. Both Latin and Cyrillic have 30 letters each; Latin has 27 monographs and three digraphs

Tarik Ćušić
University of Sarajevo, Bosnia and Herzegovina, tarik.cusic@izj.unsa.ba

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_6

(dž, lj, nj), and Cyrillic has 30 monographs. In the past, the Bosnian language was also recorded with Glagolitic, Bosnian Cyrillic (Bosančica) and Arebica.

According to the morphological classification, the Bosnian language belongs to the group of synthetic languages of the inflectional type: it has a larger number of inflections, i. e., different grammatical forms of words; it is characterised by the frequent merging of different morphemes, by a multitude of changes within individual forms and at the boundaries of morphemes, etc.

The Bosnian language belongs to the group of languages marked by the syntactic structure of SVO: Subject–Verb–Object, e. g., *Mahir sluša rok* [Mahir listens to rock.]. There are three types of word order in the Bosnian language: basic word order (grammatical-semantic), actualised word order (contextually conditioned) and obligatory word order (prosodically conditioned) (Jahić et al. 2000, p. 465-473).

In January 2021, 3.27 million people lived in Bosnia and Herzegovina (49.2% of them in urban areas): the total number of mobile connections was 3.73 million, which is 113.9% of the total population; there were 2.32 million internet users (71% of the population) and 1.8 million active social media users (55% of the population).[1]

There are more than 25,000 .ba domains registered.[2] The languages of websites under the .ba domain are mostly Bosnian, Croatian and Serbian, while some websites, due to their character and purpose, are bilingual: Bosnian – English, Croatian – English, Serbian – English and the like.

2 Technologies and Resources for Bosnian

Very few resources (i. e., corpora, language models or lexica) are available for Bosnian to date. In fact, Bosnian lacks a reference monolingual corpus that would be a valuable asset for both linguistic research and LT development. With regard to bi- or multilingual corpora, although they are rare, Bosnian is included as part of some corpora. Examples are the SETimes corpus, a parallel corpus in ten languages with its Bosnian part consisting of 2.2 million words, and the Oslo Corpus of Bosnian Texts, a 1.5 million words corpus consisting of different genres of texts published in the 1990s. The Bosnian part of the CC-100 corpus comprises 14 million tokens (Conneau et al. 2020).

In a relatively recent project aiming at compiling Web corpora of Bosnian (bsWaC) (Ljubešić and Klubička 2014), 8,388 seed URLs for Bosnian were obtained via the Google Search API queried with bigrams of mid-frequency terms obtained from corpora built with focused crawls of newspaper sites. Each TLD was crawled for 21 days with 16 cores used for document processing. The web corpus of the Bosnian language comprises 722 million tokens (Ljubešić and Klubička 2016).

[1] https://datareportal.com

[2] https://www.domaintools.com

With respect to available language technologies, Bosnian is supported in a number of machine translation systems, mainly commercial ones, like Apptek, Tradukka and iTranslate. Google Translate also supports Bosnian.

CroNER is a tool for recognising and classifying named entities in natural language texts in Croatian. CroNER recognises nine different classes of named entities. Although developed for Croatian, CroNER can successfully be applied to texts in closely related languages such as the Bosnian language.

A relatively recent (2017) mobile application for *The orthography of the Bosnian language* (Halilović 1996) can be used to learn the spelling of the Bosnian language and certain grammar rules. The mobile application allows you to search words or book chapters that contain this "orthography". This medium also allows for more flexibility than a book: You can consult "orthography" almost always, on the tram, in a cafe, during a walk. The aim was to bring the book closer to the younger generation and to promote the use of technology in education.

The Language Institute of the University of Sarajevo has developed a digital platform for the Bosnian language, e-bosanski.[3] Its goal is to offer language material about Bosnian in an online format. The material currently available is the Bosnian Dictionary of Accent Variations – Sound (Online) and Converter of Alphabets.

The Dictionary of Accent Doublets is a dictionary entry in the Bosnian Accent Manual (with a sound accent book) by a group of authors: Jasmin Hodžić, Aida Kršo and Haris Ćatović.[4] The corpus of audio recordings is designed to acquire competencies in accentuation, especially for practising general mutual accent differences in individual accents, regardless of the realised examples in everyday speech or in the Bosnian accent norm. It contains over 1,000 accent doublets selected from over 7,000 examples that make up the already excerpted material for a future study on the sources of Bosnian accentuation. Practically, this means that sound recordings for different accent variations of the same words are hosted on this platform. The Sounded Dictionary of Names is a separate part of the dictionary appendix of the future study of the Prosodem variant of personal names by the author Jasmin Hodžić. 111 names with accent variations are currently provided, i. e., recordings of different accent variations of the same names. The platform also encompasses the Accent Reader[5] and Accent Exercises.[6] The Accent Reader provides material from a hundred accented and sounded literary texts. The texts are related to everyday Bosnian life and tradition. Videos with the pronunciation of all vowels under different accents in the Bosnian language are available, including short-descending, short-ascending, and long-descending and long-ascending accents.

The platform additionally provides a Converter of Alphabets, i. e., a converter from the Latin alphabet to Glagolitic, Bosnian Cyrillic (Bosančica) and Arebica.

The Language Institute of the University of Sarajevo plans to create a large historical online dictionary of the Bosnian language that will include language material

[3] https://www.e-bosanski.ba

[4] https://www.e-bosanski.ba/rad/

[5] https://www.youtube.com/playlist?list=PL230XGW7TwJoq3ZNvg7IF7VpcsieCLW-n

[6] https://www.youtube.com/playlist?list=PL230XGW7TwJo2MgihumhTIX52_QxFBQrT

from the Middle Ages (inscriptions and charters), aljamiado texts, texts from oral literature and so-called Krajina letters. The online dictionary will provide word search functionalities, retrieving the context of the word (sentence, verse, document) from the original work.

3 Recommendations and Next Steps

As is evident from the analysis above, there are no large monolingual corpora that are representative of the modern use of the Bosnian language, or for the development of large language models (Ćušić 2022). Therefore, it is necessary to start from scratch. Current data is not sufficient in either the general or specific domains. At the national level, the Council of Ministers of Bosnia and Herzegovina is a public body that could pass the necessary acts to support the development of LT for the Bosnian language, but it is unlikely that this will happen, because language is a sensitive issue in Bosnia.

References

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proc. of the 58th Annual Meeting of the Assoc. for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. ACL, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

Ćušić, Tarik (2022). *Deliverable D1.36 Report on the Bosnian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-bosnian.pdf.

Halilović, Senahid (1996). *Pravopis bosanskoga jezika*. Preporod.

Jahić, Dževad, Senahid Halilović, and Ismail Palić (2000). *Gramatika bosanskoga jezika*. Dom štampe.

Ljubešić, Nikola and Filip Klubička (2014). "{bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian". In: *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pp. 29–35.

Ljubešić, Nikola and Filip Klubička (2016). *Bosnian web corpus bsWaC 1.1*. Jožef Stefan Institute, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1062.