EUROPEAN LANGUAGE EQUALITY

D1.36

Report on the Bosnian Language

Author	Tarik Ćušić
Dissemination level	Public
Date	28-02-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D1.36
Deliverable title	Report on the Bosnian Language
Type	Report
Number of pages	18
Status and version	Final (*Note: this document is not a contractual ELE deliverable.*)
Dissemination level	Public
Date of delivery	28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Author	Tarik Ćušić
Reviewers	Maria Giagkou, Annika Grützner-Zahn
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland
	Prof. Dr. Andy Way – andy.way@adaptcentre.ie
	European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany
	Prof. Dr. Georg Rehm – georg.rehm@dfki.de
	http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
bsWaC	Web corpora of Bosnian
CL	Computational Linguistics
DLE	Digital Language Equality
DUJIT	Society for Language Technologies
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
EU	European Union
GPU	Graphics Processing Unit
HCI	Human Computer Interaction
HPC	High-Performance Computing
LR	Language Resources/Resources
LT	Language Technology/Technologies
ML	Machine Learning
NLP	Natural Language Processing
NLU	Natural Language Understanding
NLG	Natural Language Generation
SR	Speaker Recognition
SVO	Subject-Verb-Object
TLD	Top-Level Domain

Abstract

The Bosnian language belongs to the West-South Slavic subgroup of the Slavic branch of the great Indo-European linguistic family. About 2.5 million speakers speak Bosnian as their mother tongue in Europe. Bosnian is the official language in Bosnia and Herzegovina, along with Croatian and Serbian, where it is spoken by 1.87 million people or 53%. 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 in Turkey by 100,000 to 200,000 people.

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 3 digraphs (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 data from the website datareportal.com, 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 113.9% of the total population; there were 71% internet users of the total population; there were 55% active social media users of the total population.

When it comes to the technological support for the Bosnian language, 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 research and development of language technology products have been initiated. Therefore, 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). Due to all of the above, it is imperative that the national and regional public institutions of Bosnia and Herzegovina, regardless of the level of government, initiate a program to support and fund the the development of language resources and Language Technology for Bosnian, to ensure its presence in the European linguistically diverse landscape. Such a programme should primarily focus on the construction of large Bosnian corpora, linguistic analysis tools at all linguistic levels, from grammar to semantics, machine translation, speech technologies, i. e., speech synthesis and recognition.

A small but encouraging step forward in this sense, which is not the result of a national strategy or direction, was made by the Institute of Language of the University of Sarajevo by launching the website e-bosanski, with the aim of digitising material on the Bosnian language and in the Bosnian language. In this regard, the Institute has set up an online “Bosnian Dictionary of Accent Variations” and created a digitally available writing system converter from Latin to Glagolitic, Bosančica and Arebica – three alphabets in which the Bosnian language has been recorded in the past.

Sažetak

Bosanski jezik pripada zapadnoj južnoslavenskoj podgrupi slavenske grane velike indoevropske jezičke porodice. Bosanskim kao maternjim jezikom govori oko 2,5 miliona govornika u Evropi. Bosanski je službeni jezik u Bosni i Hercegovini, uz hrvatski i srpski, gdje ga govori 1,87 miliona ljudi ili 53% i maternji je jezik Bošnjaka ali i pripadnika drugih etničkih grupa. Usto, bosanski jezik je jedan od službenih jezika u Crnoj Gori. Ovaj je jezik prepoznat kao manjinski jezik u Hrvatskoj, Srbiji, Sjevernoj Makedoniji i na Kosovu. U zapadnoj Evropi i

sjevernoj Americi bosanskim se jezikom služi oko 150.000 ljudi, a u Turskoj između 100 i 200 hiljada ljudi.

U Bosni i Hercegovini još uvijek nije donesen jedinstveni zakon o jeziku, ali se u zakonskim i podzakonskim aktima različitih nivoa vlasti u Bosni i Hercegovini uređuje pitanje zvanične upotrebe jezika. Ovi su akti uglavnom vezani sa školstvo i obrazovanje, tako da je bosanski jezik u upotrebi u školstvu i obrazovanju u Bosni i Hercegovini.

U bosanskome jeziku koriste se dva pisma: latinica i ćirilica. I latinica i ćirilica imaju po 30 slova; latinica ima 27 jednoslova i tri dvoslova (dž, lj, nj), a ćirilica 30 jednoslova. Latinica se u Bosni počinje intenzivnije upotrebljavati nakon dolaska Austro-Ugarske Monarhije. Koncept latiničnog pisma preuzet je sa Zapada, a poseban utjecaj u ortografiji ostavila je češka latinična tradicija. Uzimanje latinice za standardno pismo u direktnoj je vezi s ulaskom Bosne i Hercegovine u evropski kulturno-civilizacijski prostor, zbog čega će latinica, uz ćirilicu, postati najvažnije bosansko pismo. Kao standardizirano pismo ćirilica je u upotrebi u drugoj polovini 19. stoljeća, zahvaljujući tome što otada osmanske vlasti u Bosanskom vilajetu uvode službenu upotrebu narodnog jezika koji se imenuje bosanskim, te se počinje pisati reformiranom ćirilicom i fonetskim pravopisom. Naročito je značajna bila 1866. godina, kada se u Sarajevu otvara Vilajetska štamparija, zbog čega se ta godina uzima kao početni period upotrebe standardnog jezika pisanog fonetskim pravopisom i ćirilicom. Otada do danas savremena ćirilica ima status službenog pisma u Bosni i Hercegovini. U prošlosti je bosanski jezik bilježen i glagoljicom, bosančicom i arebicom.

Kada je riječ o morfološkoj klasifikaciji jezika, bosanski jezik pripada sintetičkim jezicima flektivnog tipa – posjeduje veći broj fleksija, tj. različitih gramatičkih oblika "istih" riječi; odlikuje se čestim i teško predvidim sjedinjavanjima različitih morfema, potom mnoštvom promjena unutar pojedinih oblika te na granicama morfema i dr.

Bosanski jezik pripada grupi jezika obilježenih sintaksičkom strukturom SVO: subjekat – verb (glagol) – objekat, npr. Mahir sluša rok. U bosanskome jeziku razlikuju se tri vrste reda riječi: osnovni red riječi (gramatičko-semantički), aktuelizirani red riječi (kontekstualno uključen) i obavezni red riječi (prozodijski uvjetovan).

Prema podacima s veb-stranice datareportal.com, u januaru 2021. godine u Bosni i Hercegovini živjelo je 3,27 miliona ljudi (49,2 u urbanim sredinama): ukupan broj mobilnih internetskih konekcija iznosio je 3,73 miliona, što je 113,9% u odnosu na ukupnu populaciju; internetskih korisnika bilo je 2,32 miliona, što obuhvata 71% ukupne populacije; aktivnih korisnika društvenih mreža bilo je 1,8 miliona ili 55% ukupne populacije.

Kada je riječ o jezičkotehnološkoj razvijenosti proizvoda za bosanski jezik, objektivno je konstatirati da ne postoje jezičke tehnologije za bosanski jezik i digitalizaciju bosanskoga jezika. Zbog toga je neophodno napraviti početne korake u jezičkotehnološkoj podršci za bosanski jezik, kako bi se spriječilo i onemogućilo njegovo digitalno izumiranje. U Bosni i Hercegovini nisu izrađeni programi usmjereni ka istraživanju i razvoju jezičkotehnoloških proizvoda u tom cilju. Stoga je bosanski jezik prisutan u digitalnoj sferi manje-više onoliko koliko je uključen u strane, višejezičke alate i resurse, koji se u najvećoj mjeri odnose na mašinsko prevođenje (Google Translate i drugi). Zbog svega navedenog nameće se imperativ da institucije Bosne i Hercegovine, bez obzira na to o kojem je nivou vlasti riječ, naprave program podrške i finansiranja izgradnje jezičkotehnoloških alata i resursa usredotočen na tehnološki potpomognutu komunikaciju putem bosanskog jezika u okviru evropske višejezičke raznolikosti. Jedan takav program uključivao bi velike bosanske korpuse, gramatičko-semantičku analizu, mašinsko prevođenje, prepoznavanje govora i sl.

Mali ali ohrabrujući iskorak u ovome smislu, koji nije rezultat nacionalne strategije ili usmjerenja, napravio je Institut za jezik Univerziteta u Sarajevu pokretanjem veb-stranice e-bosanski, s ciljem da digitalizira građu o bosanskom jeziku i na bosanskom jeziku. U tom pogledu Institut je u online-formatu postavio ozvučeni "Bosanski rječnik akcenatskih varijacija", te je napravio digitalno dostupan grafijski konverter sa latinice na glagoljicu, bosančicu i arebicu – tri pisma na kojima je bosanski jezik bilježen u prošlosti.

1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. and it seeks to not only delineate the current state of affairs for each of the European languages covered in this series, but to additionally – and most importantly – to identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

To this end, more than 40 research partners, experts in more than 30 European languages have conducted an enormous and exhaustive data collection procedure that provided a detailed, empirical and dynamic map of technology support for our languages.[1]

The report has been developed in the frame of the European Language Equality (ELE) project. With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initiatives, the ELE project develops a strategic research, innovation, and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030. Although the University of Sarajevo is not officially an ELE project partner, it recognises the importance of the ELE initiative and, through its Language Institute, acts as a collaborating partner to improve the Bosnian language presence in the digital sphere.

2 The Bosnian Language in the Digital Age

2.1 General Facts

The Bosnian language belongs to the West-South Slavic subgroup of the Slavic branch of the great Indo-European linguistic family. About 2.5 million speakers speak Bosnian as their mother tongue in Europe.[2] Bosnian is the official language in Bosnia and Herzegovina, along with Croatian and Serbian, where it is spoken by 1.87 million people or 53%. 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 3 digraphs (dž, lj, nj), and Cyrillic has 30 monographs. The Latin alphabet began to be used more intensively in Bosnia and Herzegovina after the arrival of the Austro-Hungarian Monarchy. The concept of the Latin alphabet was taken over from the West, and a special influence in orthography was left by the Czech Latin tradition. Taking the Latin alphabet as the standard alphabet is directly related to the entry of Bosnia and Herzegovina into the European cultural space, which is why the Latin alphabet, along with the Cyrillic one, have become the most important Bosnian

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they can be discovered, browsed and further investigated by means of comparative visualisations across languages.

[2] According to https://european-language-equality.eu/languages/, with data derived from Ethnologue.

alphabets. Cyrillic as a standardised alphabet was in use in the second half of the 19th century, because since then the Ottoman authorities in the Bosnian vilayet have introduced the official use of the vernacular, which is called Bosnian, and began to be written in Reformed Cyrillic and phonetic orthography. The year 1866 was especially significant, when the Vilayet printing house was opened in Sarajevo, which is why that year is considered to be the beginning of the use of the standard language written in phonetic orthography and Cyrillic. From then until today, modern Cyrillic has the status of the official alphabet in Bosnia and Herzegovina. 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 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).

2.2 Bosnian in the Digital Sphere

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, which is 71% of the total population; there were 1.8 million active social media users, which is 55% of the total population.[3]

There are more than 25,000 .ba domains registered[4]. 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.

3 What is Language Technology?

Natural language[5] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

The computational processing of human languages has been established as a specialised field known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or, more generally, Language Technology (LT). While there are differences in focus and orientation, since CL is more informed by linguistics and NLP by computer science, LT is a more neutral term. In fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably AI), mathematics and psychology among others. In practice, these commu-

[3] Source: https://datareportal.com, January 2021
[4] https://www.domaintools.com, December 2021
[5] This section has been provided by the editors. It is an adapted summary of Agerri et al. (2021) and of Sections 1 and 2 of Aldabe et al. (2021).

nities work closely together, combining methods and approaches inspired by both, together making up *language-centric AI*.

Language Technology is the multidisciplinary scientific and technological field that is concerned with studying and developing systems capable of processing, analysing, producing and understanding human languages, whether they are written, spoken or embodied.

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar(Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in machine learning, rule-based systems have been displaced by data-based ones, i. e., systems that learn implicitly from examples. In the recent decade of 2010s we observed a radical technological change in NLP: the use of multilayer neural networks able to solve various sequential labelling problems. The success of this approach lies in the ability of neural networks to learn continuous vector representations of the words (or word embeddings) using vast amounts of unlabelled data and using only some labelled data for fine-tuning.

In recent years, the LT community has been witnessing the emergence of powerful new deep learning techniques and tools that are revolutionising the way in which LT tasks are approached. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. The success in these areas of AI has been possible because of the conjunction of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual data), 3) increase in high performance computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches.

LT is trying to provide solutions for the following main application areas:

- **Text Analysis** which aims at identifying and labelling the linguistic information underlying any text in natural language. This includes the recognition of word, phrase, sentence and section boundaries, recognition of morphological features of words, of syntactic and semantic roles as well as capturing the relations that link text constituents together.
- **Speech processing** aims at allowing humans to communicate with electronic devices through voice. Some of the main areas in Speech Technology are Text to Speech Synthesis, i. e., the generation of speech given a piece of text, Automatic Speech Recognition, i. e., the conversion of speech signal into text, and Speaker Recognition (SR).
- **Machine Translation**, i. e., the automatic translation from one natural language into another.
- **Information Extraction and Information Retrieval** which aim at extracting structured information from unstructured documents, finding appropriate pieces of information in large collections of unstructured material, such as the internet, and providing the documents or text snippets that include the answer to a user's query.
- **Natural Language Generation (NLG)**. NLG is the task of automatically generating texts. Summarisation, i. e., the generation of a summary, the generation of paraphrases, text re-writing, simplification and generation of questions are some example applications of NLG.

- **Human-Computer Interaction** which aims at developing systems that allow the user to converse with computers using natural language (text, speech and non-verbal communication signals, such as gestures and facial expressions). A very popular application within this area are conversational agents (better known as chatbots).

LT is already fused in our everyday lives. As individual users we may be using it without even realising it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name just very few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in *educational* settings and applications, for instance for educational content mining, for the automatic assessment of free text answers, for providing feedback to learners and teachers, for the evaluation of pronunciation in a foreign language and much more. In the *law/legal* domain, LT proves an indispensable component for several tasks, from search, classification and codification of huge legal databases to legal question answering and prediction of court decisions.

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[6]

4 Language Technology for Bosnian

4.1 Language Data and Tools

Very few language 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 Language Technology development. With regard to bi- or multilingual corpora, although they are rare, Bosnian is included as part of some corpora. Examples of such bilingual or multilingual corpora 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, which is a 1.5 million words corpus consisting of different genres of texts that were 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. Those terms were obtained from corpora that were built with focused crawls of newspaper sites as part of our previous research. Each TLD was crawled for 21 days with 16 cores used for document processing. According to Ljubešić and Klubička (2016), the web corpus of the Bosnian language comprises 722 million tokens.

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

[6] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at an annual rate of 18.4% from 2020 to 2028 (https://www.globenewswire.com/news-release/2021/03/22/2196622/0/en/Global-Natural-Language-Processing-Market-to-Grow-at-a-CAGR-of-18-4-from-2020-to-2028.html). A different report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at USD 13 billion in the year 2020 and is projected to reach USD 25.7 billion by 2027, growing at an annual rate of 10.3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

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 an electronic platform for the Bosnian language, e-bosanski.[7] The goal of this platform is to offer language material about Bosnian in an online format. The language 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.[8] The corpus of audio recordings is designed to acquire competencies in accentuation, especially for practicing 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 (thousand) 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 provided here currently and this list will be expanded. Practically, that means that sound recordings for different accent variations of the same names are hosted on this platform. The platform also encompasses the Accent Reader[9] and Accent Exercises.[10] The Accent Reader provides material from a hundred accented and sounded literary texts. The texts in the Reader are related to everyday Bosnian life and tradition. Videos with the pronunciation of all vowels under different accents in the Bosnian language were 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 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.

As 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 construction of massive language models. Therefore, it is necessary to start from scratch. Current data is not sufficient in either the general or specific domains. There are no collections of texts on social media in the context of hate speech, propaganda, or fake news detection. There are no bilingual resources that may be of particular interest for social, economic, or other reasons in Bosnia and Herzegovina. In the context of LT applications for the Bosnian language, there are no tools/services for basic NLP tasks. There is no application made for speech recognition in the Bosnian language.

[7] https://www.e-bosanski.ba
[8] https://www.e-bosanski.ba/rad/
[9] https://www.youtube.com/playlist?list=PL230XGW7TwJoq3ZNvg7IF7VpcsieCLW-n
[10] https://www.youtube.com/playlist?list=PL230XGW7TwJo2MgihumhTIX52_QxFBQrT

4.2 Projects, Initiatives, Stakeholders

There is no national program for Language Technology or Artificial Intelligence. There are no organisations in Bosnia that research language technologies. Earlier, until 2011, there was the Society for Language Technologies (DUJIT), which was engaged in research on language technology. Unfortunately, there has been no social or institutional interest in continuing the work of this Society. There is no public body in Bosnia that deals with Bosnian language policy. There were ideas to form a council for the standardisation of the Bosnian language, but everything remained on the ideas. 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 LT and AI for the Bosnian language, but it is unlikely that this will happen, because language is a sensitive issue in Bosnia.

5 Cross-Language Comparison

The LT field[11] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded results unforeseeable before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[12] broadly categorised into a number of core LT application areas:
  - Text processing (e. g. part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g. search and information mining)
  - Translation technologies (e. g. machine translation, computer-aided translation)
  - Natural language generation (e. g. text summarisation, simplification)
  - Speech processing (e. g. speech synthesis, speech recognition)
  - Image/video processing (e. g. facial expression recognition)
  - Human-computer interaction (e. g. tools for conversational systems)
- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora

[11] This section has been provided by the editors.

[12] Tools tagged as “language independent” without mentioning any specific language are not taken into account. Such tools can certainly be applied to a number of languages, either as readily applicable or following fine-tuning, adaptation, training on language-specific data etc., yet their exact language coverage or readiness is difficult to ascertain.

- Multimodal corpora (incl. speech, image, video)
- Models
- Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

5.2 Levels of Technology Support

We measured the relative technology support for 87 national, regional and minority European languages with regard to each of the dimensions mentioned above based on their respective coverage in the ELG catalogue. For the types of resources and application areas, the respective percentage of resources that support a specific language over the total number of resources of the same type was calculated, as well as their average. Subsequently each language was assigned to one band per resource type and per application area and to an overall band, on a four-point scale, inspired by the scale used in the META-NET White Paper Series, as follows:

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[13]

The overall level of support for a language was calculated based on the average coverage in all dimensions investigated.

5.3 European Language Grid as Ground Truth

At the time of writing (January 2022), the ELG catalogue comprises more than 11,500 metadata records, encompassing both data and tools/services, covering almost all European languages – both official and regional/minority ones. The ELG platform harvests several major LR/LT repositories[14] and, on top of that, more than 6,000 additional language resources and tools were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the Digital Language Equality (DLE) metric will allow for dynamic analyses and

[13] The thresholds for defining the four bands were informed by an exploratory k-means 4-cluster analysis based on all data per application and resource type, in order to investigate the boundaries of naturally occurring clusters in the data. The boundaries of the clusters (i. e., 3%, 10% and 30%) were then used to define the bands per application area and resource type.

[14] At the time of writing, ELG harvests ELRC-SHARE, LINDAT/CLARIAH-CZ, CLARIN.SI, CLARIN-PL and HuggingFace.

calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[15]

For the purposes of high-level comparison in this report, the results presented here are based on relative counts of entries in the ELG for the varying types of data resources and tools/services for each language. As such, the positioning of each language into a specific level of technology support is subject to change and it reflects a snapshot of the available resources on January 2022.

That said, we consider the current status of the ELG repository and the higher level findings below adequately representative with regard to the current existence of LT resources for Europe's languages.

5.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 1 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g. German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[16] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

While a fifth level, *excellent support*, could have been foreseen in addition to the four levels described in Section 5.2, we decided not to consider this level for the grouping of languages. Currently no natural language is optimally supported by technology, i. e., the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning, emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural language Understanding has been reached.

The results of the present comparative evaluation reflect, in terms of distribution and imbalance, the results of the META-NET White Paper Series (Rehm and Uszkoreit, 2012). The complexities of the analyses clearly differ across 2012 and 2022 studies, and as such, a di-

[15] Interactive comparison visualisations of the technology support of Europe's languages will be possible on the ELG website using a dedicated dashboard, which dynamically analyses the resources available in the ELG repository, from the middle of 2022 onwards.

[16] In addition to the languages listed in Table 1, ELE also investigated Alsatian, Aragonese, Arberesh, Aromanian, Asturian, Breton, Cimbrian, Continental Southern Italian (Neapolitan), Cornish, Eastern Frisian, Emilian, FrancoProvencal (Arpitan), Friulian, Gallo, Griko, Inari Sami, Karelian, Kashubian, Ladin, Latgalian, Ligurian, Lombard, Lower Sorbian, Lule Sami, Mocheno, Northern Frisian, Northern Sami, Picard, Piedmontese, Pite Sami, Romagnol, Romany, Rusyn, Sardinian, Scottish Gaelic, Sicilian, Skolt Sami, Southern Sami, Tatar, Tornedalian Finnish, Venetian, Võro, Walser, Yiddish.

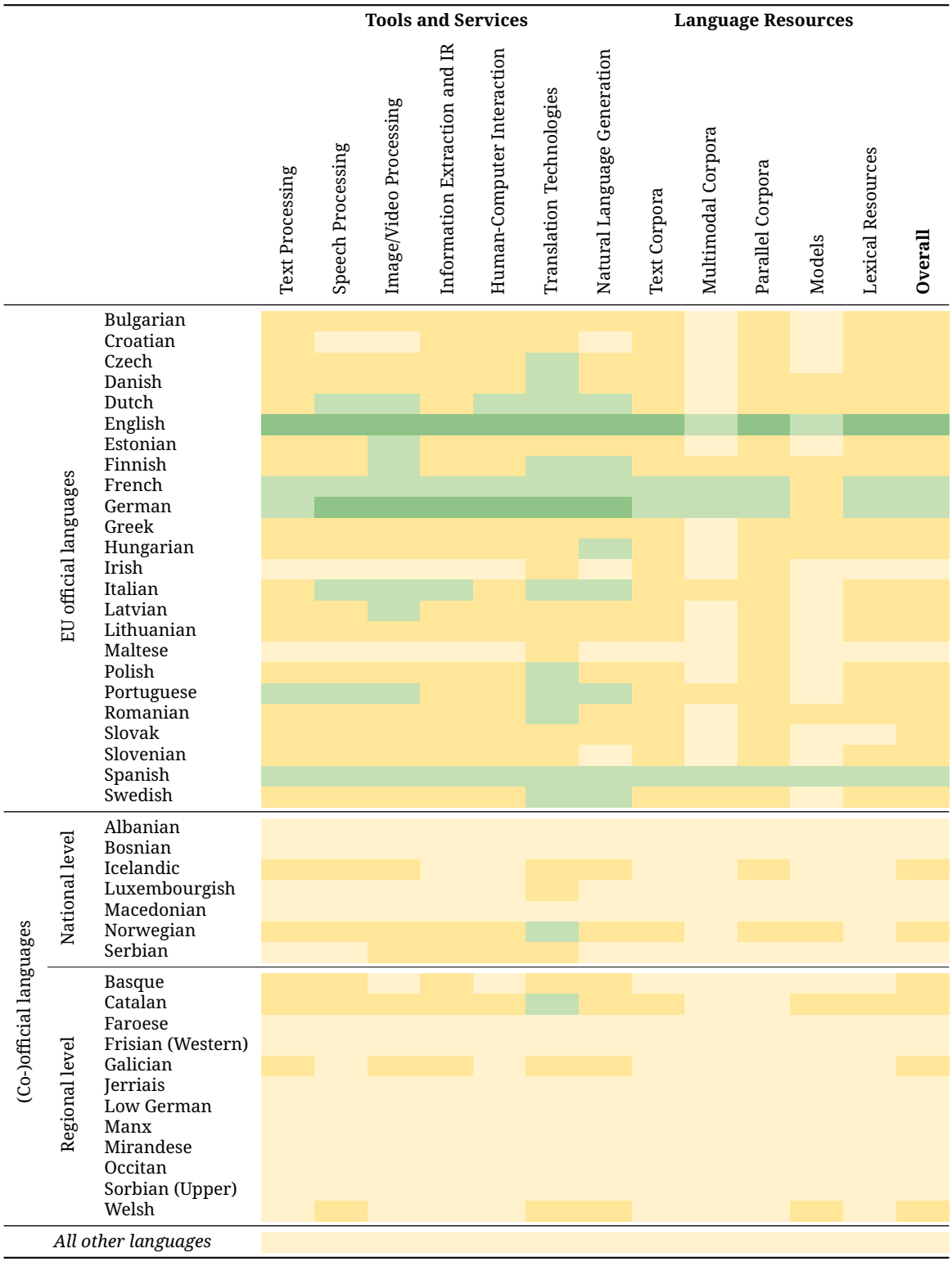

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

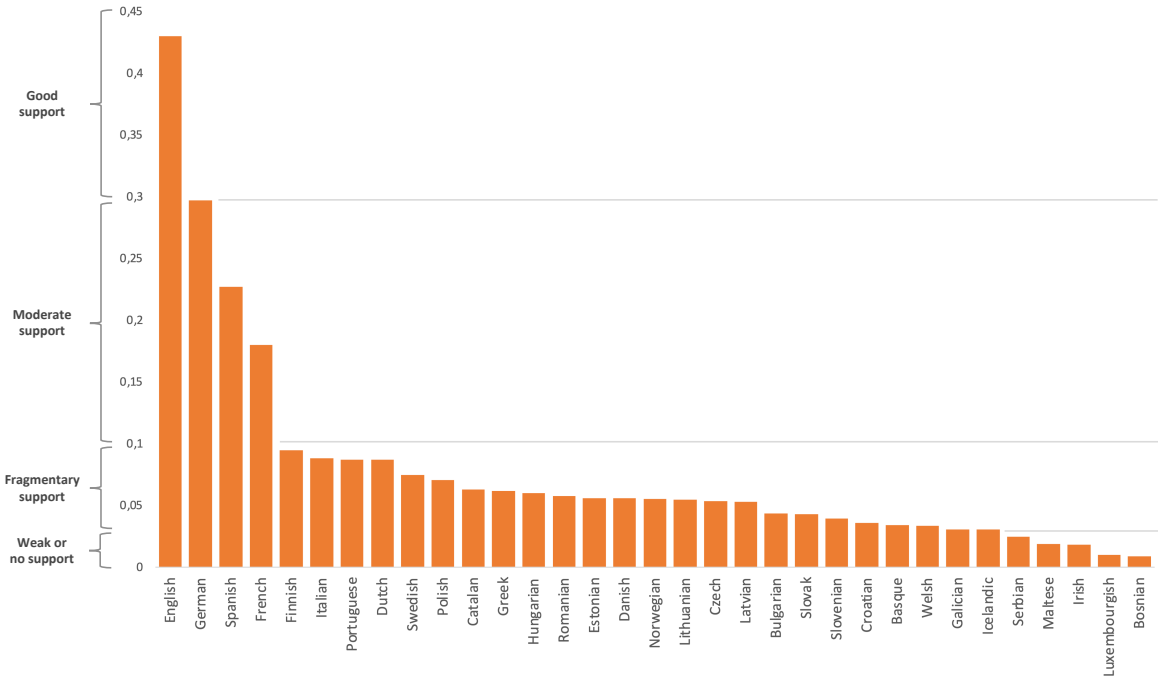

Figure 1: Overall state of technology support for selected European languages (2022)

rect comparison between the two studies can therefore not be made. However, we can instead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e. g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

The results of this analysis are only informative of the relative positioning of languages, but not of the progress achieved within a specific language. The LT field as a whole has significantly progressed in the last ten years and remarkable progress has been achieved for specific languages in terms of quantity, quality and coverage of tools and language resources. Yet, the abysmal distance between the best supported languages and the minimally supported ones is still evidenced in 2022. It is exactly this distance that needs to be ideally eliminated, if not at least reduced, in order to move towards Digital Language Equality and avert the risks of digital extinction.

6 Summary and Conclusions

From the analysis of the availability of language resources and technologies, it can be concluded that there are large gaps and shortcomings in terms of language technologies for the Bosnian language. Apart from some language technologies, e. g. machine translation, which are provided as part of some commercial service offerings (Google Translate, Apptek, Tradukka, iTranslate), there are no language technologies that support the Bosnian language. A corpus of Bosnian is also missing, either unannotated or annotated for morphological, syntactic or semantic structures. This is an objective weakness of the digital Bosnian language. Until now, language technologies for the Bosnian language have not been addressed by either national programmes in Bosnia and Herzegovina or by individuals.

A positive fist step has been achieved by the Language Institute of the University of Sarajevo, which seeks to digitise language material and build language technologies that could serve a wider audience but also to generally enhance the presence of the Bosnian language in the digital sphere.

To this end, the Language Institute is planning to develop a website (e-bosanski) dedicated to the digitisation of the Bosnian language and to digital language material. This website is in its initial development phase.

It is important to actually initiate LT research for Bosnian, so that Bosnian tries to keep pace with other European languages in this field and it achieves a certain level of digital readiness.

References

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. European Language Equality D1.2: Report on the state of the art in Language Technology and Language-centric AI, September 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf.

Itziar Aldabe, Georg Rehm, German Rigau, and Andy Way. European Language Equality D3.1: Report on existing strategic documents and projects in LT/AI, November 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf.

Noam Chomsky. *Syntactic structures*. The Hague: Mouton, 1957.

A Conneau, K Khandelwal, N Goyal, V Chaudhary, G Wenzek, F Guzmán, E Grave, M Ott, L Zettlemoyer, and V Stoyanov. Unsupervised cross-lingual representation learning at scale. arxiv 2019. *arXiv preprint arXiv:1911.02116*, 2020.

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

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

Nikola Ljubešić and Filip Klubička. {bs, hr, sr} wac-web corpora of bosnian, croatian and serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, 2014.

Nikola Ljubešić and Filip Klubička. Bosnian web corpus bswac 1.1. 2016.

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc., 2012. Springer.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.