EUROPEAN LANGUAGE EQUALITY

D1.35

Report on the Serbian Language

Authors	Cvetana Krstev, Ranka Stanković
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	D1.35
Deliverable title	Report on the Serbian Language
Type	Report
Number of pages	25
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
Authors	Cvetana Krstev, Ranka Stanković
Reviewers	Stefanie Hegele, Maria Eskevich
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
ASR	Automatic Speech Recognition
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CLASSLA	CLARIN Knowledge Centre for South Slavic languages
CQP	Corpus Query Processor
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
ELTeC	European Literary Text Collection
GPU	Graphics Processing Unit
HCI	Human Computer Interaction (see HMI)
HMI	Human Machine Interaction (see HCI)
HPC	High-Performance Computing
ICT	Information and Communication Technologies
JeRTeh	Society of Language Technologies and Resources (Serbia)
LR	Language Resources/Resources
LT	Language Technology/Technologies
ML	Machine Learning
MT	Machine Translation
NE	Named Entity
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
POS	Part-Of-Speech
SR	Speaker Recognition
SrpMD	Serbian Morphological Dictionaries
TEI	Text Encoding Initiative
TTS	Text to Speech
UB	University of Belgrade
UD	Universla Dependencies

Abstract

This report presents the results of an investigation of the level of support the European languages receive through technology. The aim of this investigation is to identify factors that hinder the development of needed language technologies and languages that lack the majority of language resources, tools and applications. By identifying such weaknesses it will lay the grounds for a comprehensive, evidence-based proposal of required measures for achieving Digital Language Equality in Europe by 2030.

The focus of this deliverable is the Serbian Language and its presence in the digital world. Serbian is the official language of the Republic of Serbia, spoken as a mother tongue by 88.1% of its citizens. The number of Serbian-speaking persons living abroad is large but difficult to estimate. According to recent statistical data provided by official authorities, Serbian citizens are equipped to live in the digital world: only 17.6% of adult citizens never used computer, and 10.4% of them never used the Internet. Data for the younger population aged 16–24 is even more favorable: less than 5% of young Serbian citizens have never used a computer, and less than 2% of them have never used the Internet. 81% of households and enterprises have internet connection. However, only 19.3% of enterprises have ICT professionals among employees.

This data shows that Serbian citizens are ready to use language technologies (LT). 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. The main application areas of LT are: text analysis, speech processing, machine translation, information extraction, information retrieval, natural language generation, and human-computer interaction. LT is already fused in our everyday lives and it is an important, although sometimes invisible, ingredient of applications that cut across various sectors and domains. LT is also one of the most important AI areas with a fast growing economic impact.

The overview of LT for Serbian showed that some resources for Serbian are rich and diverse, like mono-, bi- and multi-lingual corpora and lexicons, while some types of resources are still rare (conceptual resources, models and grammars), and some practically do not exist (multimodal corpora). There exist several and robust tools for tokenisation, POS and morphosyntactic tagging, lemmatisation, and named-entity recognition. Speech technologies are well developed but exist only for commercial use, some services for information extraction and information retrieval were developed, while translation technologies, language generation, summarisation, and human-computer interaction are underdeveloped.

In recent years, the government of Serbia has recognised the importance of Artificial Intelligence, and Natural Language Processing as one of the important AI applications. Having in mind that in the past, as well as today, researchers working on LT for Serbian are mostly affiliated with state universities, those institutions require stable and adequate funding. However, there is still no specific LT-related funding in Serbia aimed at filling the recognised gaps in LT for Serbian.

The evaluation of LT for 87 languages was performed on the basis of more than 10,000 items catalogued in the European Language Grid (as of January 2022). The current state of technology support was measured by the availability of tools and services categorised in the previously mentioned core LT technologies, as well as by the availability of resources (text and parallel corpora, multimodal corpora, language models and lexical resources) that can be used for training or evaluation. The analysis showed that the Serbian language is either weakly or fragmentarily supported with respect to all mentioned dimensions. As a result, the conclusion is that the overall support for the Serbian language is weak.

Апстракт

Овај извештај је резултат истраживања подршке коју технологија пружа европским језицима. Циљ истраживања је био да се идентификују фактори који ометају развој потребних језичких технологија и језици којима недостаје највећи број језичких ресурса, алата и апликација. Идентификовањем ових слабости поставиће се основа за свеобухватни и документован предлог мера које је неопходно предузети да би се до 2030. године остварила дигитална равноправност језика у Европи.

У фокусу овог извештаја је српски језик и његово присуство у дигиталном свету. Српски је званични језик Републике Србије и матерњи језик 88.1% њених грађана. Број људи којима је српски матерњи језик а који не живе у Србији је тешко проценити. Према најновијим званичним статистичким подацима грађани Србије су добро опремљени за живот у дигиталном свету: само 17.6% одраслих грађана никада није користило рачунар, док 10.4% њих никада није користило Интернет. Подаци за млађу популацију узраста 16–24 година су још убедљивији: мање од 5% младих у Србији никада није користило рачунар, док мање од 2% њих никада није користило Интернет. Интернет везу поседује 81% домаћинстава и сва предузећа. Ипак, само 19.3% предузећа запошљава ИТ стручњаке.

Ови подаци говоре да су грађани Србије спремни да користе језичке технологије. Језичке технологије су мултидисциплинарна научна и технолошка област у оквиру које се изучавају и развијају системи који могу да обрађују, анализирају, производе и разумеју језике којима се људи служе, било да су у писаном, говорном или знаковном облику. Најважнија поља примене језичких технологија су: анализа текста, обрада говора, машинско превођење, екстракција информација, проналажење информација, генерисање природног језика и интеракција између човека и рачунара. Језичке технологије су већ укључене у свакодневни живот људи и представљају важан, иако понекад невидљив, део разноврсних апликација које се користе у различитим доменима. Такође, језичке технологије представљају једну од најзначајнијих подобласти вештачке интелигенције чији економски утицај брзо расте.

Преглед језичких технологија за српски језик је показао да су неки ресурси за српски богати и разноврсни. То пре свега важи за једнојезичне, двојезичне и вишејезичне корпусе и за лексиконе и речнике. Неки ресурси су још увек ретки, попут концептуалних ресурса и граматика, док неки, као мултимодални корпуси, практично не постоје. Постоји више робустних алата за токенизацију, морфосинтаксично тагирање и тагирање врстом речи, лематизацију и препознавање именованих ентитета. Језичке технологије су добро развијене али су оне доступне углавном само комерцијално. Развијено је више сервиса за проналажење и екстракцију информација док су технологије намењене аутоматском превођењу, генерисању текста, сумаризацији и интеракцији човека и рачунара још увек недовољно развијене.

Последњих година Влада Републике Србије препознаје значај вештачке интелигенције и обраде природних језика, као једне њене важне примене. Имајући у виду да истраживачи који су радили а и данас раде у области језичких технологија потичу углавном са државних универзитета, овим институција је потребно стабилно и адекватно финансирање. Међутим, још увек не постоји линија финансирања намењена посебно језичким технологијама, било у оквиру вештачке интелигенције или не (на пример, за развој граматика за српски језик неопходних за развој језичких технологија), чији би циљ био попуњавање уочених недостатака у њиховом развоју за српски језик.

На основу више од 10.000 ставки каталогизираних до јануара 2022. године у оквиру Европске језичке мреже (European Language Grid) обављена је евалуација језичких технологија за 87 европских језика. Тренутно стање технолошке подршке мерено је

доступношћу алата и сервиса категорисаних у претходно поменуте основне језичке технологије, као и доступношћу ресурса (једнојезични и паралелни корпуси, мултимодални корпуси, језички модели и лексички ресурси) који се могу користити за тренирање или евалуацију. Ова анализа је показала да је српски језик слабо или делимично подржан према свим поменутим димензијама, а одатле следи закључак да је укупна подршка језичким ресурсима српском језику слаба.

Ово истраживање је омогућило да се упореде језичке технологије за српски са онима које постоје за енглески и друге језике са добром технолошком подршком. Иако се у протеклих десет година уочава напредак у развоју језичких технологија за српски језик чији резултат су бољи и разноврснији ресурси и алати, цела област је у том периоду веома напредовала што је донело још више напредних ресурса и алата за енглески и друге језике који су већ ионако предњачили. С тога се данас може приметити да се упркос напретку који је постигнут за српски језик разлика између српског и других добро подржаних језика није смањила. Ово истраживање је такође указало да је разлика између језика који су блиски српском просторно, историјски и по броју говорника, као што су бугарски, словеначки и хрватски, и језика који су релативно добро подржани мања, премда још увек знатна. Политике које ове земље спроводе да би промовисале и подржале језичке технологије за своје језике могу да укажу на правац који треба следити да би се унапредиле језичке технологије за српски.

Упркос вредним достигнућима која су постигнута, српски језик је и даље у неповољном положају што носи опасност да за неколико година говорник српског језика не би могао да користи добробити наступајуће револуције у области вештачке интелигенције и језичких технологија. Да би се то спречило, потребно је више финансирања и на националном и на међународном плану. На међународном плану српском и другим слабо подржаним језицима би користили пројекти за пренос знања чији циљ не би било пуко пресликавање постојећих решења за енглески језик већ који би подржавали производњу адекватних ресурса и алата за угрожене језике.

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.[2] 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.

[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] https://european-language-equality.eu

2 The Serbian Language in the Digital Age

2.1 General Facts

Standard Serbian is the national language of Serbs and the official language in the Republic of Serbia. It was formed on the basis of Ekavian and Ijekavian Neo-Štokavian South Slavic Dialects and its form was determined by the reformer of the written language of Serbs Vuk Karadžić (1787-1864), who at the same time reformed both the Cyrillic alphabet and orthography. In the 20th century, in the federal state of Yugoslavia, this language was officially encompassed by *Serbo-Croatian*, a name that implied a linguistic unity with Croats (and later with other nations whose languages were based on Neo-Štokavian dialects). In the last decade of the 20th century in Serbia the name Serbo-Croatian was replaced by the name Serbian (Popović, 2004). The Constitution of the Republic of Serbia from 2006 stipulates: "The Serbian language and the Cyrillic alphabet shell be in official use in the Republic of Serbia" (Ustav, 2006). However, the Latin alphabet is in widespread use. According to data from the Union Library Catalog of Serbia, in 2020 1,612 monographs were published in Cyrillic while 1,554 were published in the Latin alphabet.

According to the 2011 census[3] the population of Serbia is 7,186,862, and Serbian is the mother tongue of 88.1% of the population. To this number one should add the ethnic Serb population in other parts of former Yugoslavia (a number not easy to determine). The Serbian diaspora lives primarily in a number of countries of Central and Western Europe, in the USA, Canada and Australia, and their knowledge of Serbian is mainly determined by the generation of immigrants they belong to.

According to the 2011 census, the structure of minority languages spoken in Serbia is the following: Hungarian 3.4%, Bosnian 1.9%, Roma 1.4%, Slovak 0.7%, Wallach 0.6%. The remaining languages are spoken by 1.6% of the population, whereas for 2.33% of the population the data is unknown.

Translations to and from Serbian represent an important activity. During 2020 a total of 3,725 works were translated, mostly from English (46.2%), followed by French, Italian, Russian, Spanish, etc. As for translations from Serbian into other languages, 1,054 works were published in 2020, and mostly translated into English (39.4%). As for minority languages and languages in close contact, most works were translated from or to Hungarian, followed by Slovak, Macedonian, Slovene, Bulgarian, Croatian and Bosnian. Translations to and from Hungarian, Slovak, Bosnian and Croatian consist mostly of school textbooks.

2.2 Serbian in the Digital Sphere

There were 6,406,827 Internet users in Serbia on 31 December 2020, or 73.4% out of an estimated population of 8,733,407 , while there were 3,926,000 users of Facebook, or 45.1%.[4]

The number of sites with the `.rs` top domain is 121,383, the number of mail servers is 22,192, and the number of IP addresses is 2,519,545.[5]

According to the 2011 census, 34.2% out of 6,161,584 inhabitants (2,971,868 men, 3,189,716 women) 15 years old or older had computer literacy skills, while 14.8% were partially computer literate. Women fall behind men: 32.8% women had computer literally skills (vs. 35.7% men), while 14.0% of them were partially computer literates (vs. 15.6%). In order to assess

[3] Statistical Office of the Republic of Serbia – Census 2011 (https://www.stat.gov.rs/sr-latn/oblasti/popis/popis-2011); starting from 1999 census of the Republic of Serbia do not include data from Kosovo.

[4] Source: Internet in Europe (https://www.internetworldstats.com/stats4.htm); it should be noted that the estimated population by Internet World Statistics (8,733,407) differs significantly from data provided by the Statistical Office of the Republic of Serbia (7,186,862).

[5] Source: Domain Count Statistics for top level domains (https://research.domaintools.com/statistics/tld-counts/)

somebody's skills, people were asked to reply whether they knew how to use text processing, spreadsheets, e-mail and the Internet.

The Statistical Office of the Republic of Serbia collects data about the use of ICT in Serbia each year.[6] According to their incomplete data for 2021 published on October 22, 2021, the percentage of citizens between 16 and 74 years of age that have never used a computer was 17.6% (compared to 38.9% in 2012), while the percentage of citizens that never have used he Internet was 10.4% (compared to 48.4% in 2012). A computer is used regularly[7] by 74.8% of the citizens (compared to 55.9% in 2012), while the Internet is used regularly by 81.2% of the citizens (compared to 48.4% in 2012). Data for the 16-24 age group is more favorable: computers are regularly used by 95.2% of the young people, and Internet by 98.1%. Citizens with higher education (at least 2 years of regular studies after the secondary school) use computers regularly (90.1%), as well as the Internet (91.1%). Women fall behind men both in the regular use of computers (70.9% compared to 78.8% male users) and the use of Internet (78.8% compared to 83.6%). When it comes to e-commerce, 42.3% of the population regularly use it (compared to 16.6% in 2012); in this case women more often than men (45.2% vs. 39.4%).

The Statistical Office of the Republic of Serbia published complete official results of the use of ICT study for the year 2020 (ICT2020, 2020), and there one can find that 74.3% of households possessed a computer (61.8% in rural districts), while 81.0 % of households had an Internet connection (70.4% in rural districts), out of which 90.5% had a broadband connection. Internet was used for private purposes mostly to send messages – via Skype, Messinger, WhatsUp, Viber – (84.0% of users), to telephone (80.5%), to read on-line news and magazines (73.6%), to use social media – Facebook, Twitter – (71.2%),[8] and to seek information about goods and services (69.6%). As for e-government, this study showed that 37% of Internet users used Internet services instead of personally visiting public institutions and administrative bodies. These cases were used mostly to seek information on public institutions websites (34.0%), to download forms (25.2%) and to upload completed forms (23.9%).

As for business, 100% of enterprises in Serbia had Internet connection in 2020, 98.4% with a broadband connection. In 36.0% of enterprises, Internet was used by 1% to 24% employees, while in a similar number of enterprises (35.7%) the Internet was used by 75% to 100% of employees. 84.4% of enterprises have their own websites (compared to 80.8% in 2016). 27.9% enterprises sold their goods and/or services via the Internet during 2019, while 18.6% of enterprises pay via Internet cloud services. 19.3% of enterprises have ICT professionals among their employees.

3 What is Language Technology?

Natural language[9] 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,

[6] The use of ICT – official statistics (https://www.stat.gov.rs/sr-Latn/oblasti/upotreba-ikt/upotreba-ikt-pojedinci)
[7] "Regularly" is defined in this research as "at least once in last 3 months".
[8] This number differs significantly from the one provided by the Internet World Statistics – 45.1%.
[9] 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).

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 communities 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 revolutionizing 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 realizing 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.[10]

4 Language Technology for Serbian

While some resources for Serbian are rich and diverse (mono-, bi- and multi-lingual corpora and lexicons), other types of resources are rare (conceptual resources, models and grammars). Moreover, multimodal corpora are still lacking in the landscape. There exist several diverse tools for tokenisation, POS and morphosyntactic tagging, lemmatisation, and named-entity recognition (NER). Speech technologies are well developed but exist only for commercial use, some services for information extraction and information retrieval were developed, while translation technologies, language generation, summarisation, and human-computer interaction are underdeveloped.

4.1 Language Data

Mono-, bi- and multi-lingual corpora

A variety of texts and corpora for Serbian was produced and made available that vary in types and levels of annotation. The variety of corpora as well as their availability has improved significantly in recent 10 years (Vitas et al., 2012). As for text types, many of these text collections/corpora contain data obtained from various news portals (SETimes, MetaLangNEWS-sr), Wikipedia (CLASSLA), Twitter (Serbian Twitter training corpus ReLDI-NormTagNER-sr

[10] 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://tinyurl.com/4tdwtx7u). 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).

2.1), movie reviews (SerbMR-3C), or by web crawling (srWac) (Ljubešić and Klubička, 2014). Some collections/corpora are represented as raw texts, others are annotated with POS and lemmas, while a few are fully morphologically and/or NE annotated. A few collections/corpora contain transcribed texts (Serbian forms of address 1.0, Spoken Torlak dialect corpus 1.0). Some of these collections/text were prepared for a special purpose, such as sentiment analysis (collected Movie Reviews).[11] Two corpora/datasets were developed for text similarity and text paraphrasing analysis, each consisting of approx. 1,200 sentences pairs annotated for similarity, or paraphrasing (Batanović et al., 2018).

As for prepared corpora, to balance criteria with the aim of representing the language in general to a certain extent , there are two corpora available on the web, developed by researchers from the University of Belgrade (UB) and the Society of Language Resources and Tools – JeRTeh.[12] The Corpus of Contemporary Serbian *SrpKor2013*[13] contains more than 122 million tokens annotated with Part-Of-Speech and lemmas (Vitas and Krstev, 2012). For corpus management and search, the IMS/CQP software produced by the University of Stuttgart is used. It is available for search for registered users.

At the end of 2021 the new version of this corpus, dubbed SrpKor2021 was published. It contains new texts compared to SrpKor2013 comprising of more than 600 million words and it is annotated with Part-of-Speech and lemmas. The new system for the management of corpus texts will enable its permanent enhancement. Besides newspaper texts, it contains texts from Wikipedia, literary texts, textbooks on various topics, PhD theses, as well as other texts from various domains. Among literary texts are 100 novels forming the Serbian subcollection of the European Literary Text Collection (ELTeC) corpus prepared in the scope of the COST action CA16204 *Distant Reading for European Literary History.* For the corpus management and search NoSketch (open source version of Sketch Engine) is used, and it is available for registered users at the site of JeRTeh.[14] However, some of its parts are open (Wikipedia, SrpELTeC). Besides SrpKor2021, NoSketch at the site of JeRTeh suports various domain corpora (mathematics, mining, geology, etc.).

There are numerous multilingual collections/corpora that include Serbian as one of the languages. Some of them, registered at ELG, are *CLASSLA-Wikipedia 1.0*, *Plain text Wikipedia dump 2018*, *W2C – Web to Corpus* , *Text collection for training the BERTić* (Ljubešić, 2021), to mention just a few. The Serbian part of these collections can be used independently for monolingual processing. There are some multi- or bi-lingual collections/corpora in which language components are aligned, for instance *INTERA Corpus – the Serbian-English part*, *SETimes – A Parallel Corpus* and *Parallel corpus srenWaC 1.0* (see footnote 11).

The English/Serbian aligned corpus developed by UB/JeRTeh and supported by *Bibliša* digital library,[15] contains texts from various domains having more than 130,000 aligned sentences and it is being permanently enhanced (Stanković et al., 2017). Two other bilingual corpora were built by the same group that contain text of various types, among them a large part are literary texts. These are English/Serbian and French/Serbian aligned corpora containing in both languages approx. 4.5, or 1.7 respectively, million tokens, both accessible on the Web for searching. The German/Serbian corpus supported by *Bibliša* contains only literary texts, consisting of 14 novels with 48,004 aligned sentences.

Some of these multi- or bi-lingual collections/corpora were build for special purposes, such as the *MT corpus PErr 1.0* and *exams* intended for multi- and cross-lingual question answering. The *Deep Universal Dependencies 2.8* contains a collection of treebanks derived semi-automatically from Universal Dependencies (UD). In addition, multi-lingual versions of *Bible* and *1984* include also Serbian (see footnote 11).

[11] More information about resources mentioned in this paragraph can be found in the ELG catalogue.
[12] The Society of Language Resources and Tools (http://jerteh.rs)
[13] The Corpus of Contemporary Serbian (http://www.korpus.matf.bg.ac.rs/korpus/login.php)
[14] SrpKor2021 (https://noske.jerteh.rs)
[15] *Bibliša* (http://jerteh.rs/biblisha/Default.aspx)

Lexical/conceptual resources

Lexical resources for Serbian are numerous. Among the many monolingual lexicons are *srLex*, an inflectional lexicon based on the MULTEXT-East V5 tagset (Krstev et al., 2004), *KORLEX – Serbian Lexicon*, the Serbian part of the bilingual lexicon for English-Serbian and *MULTEXT-East non-commercial lexicons 4.0* (see footnote 11).

By far the most comprehensive lexical resource for Serbian that covers both simple and multi-word units, general lexica, proper names, and domain lexica, is *Serbian Morphological Dictionaries – SrpMD* developed by UB/JerTeh researchers (Krstev, 2008). The description of entries in these dictionaries goes beyond morphological descriptions. It covers, to a certain extent, semantics, usage, pronunciation, etymology, domains, derivational relations, etc. and it is being permanently updated. Currently, it contains more than 205,000 simple- and almost 23,000 multi-word units. At the end of 2021 these dictionaries became open for search through the platform *Leximirka*. Registration is obligatory for different types of users having various levels of accessibility to data. The part of this lexicon containing word forms for approx. 85,000 lemmas are made public and registered on ELG at the end of 2021.[16]

NLP resources for Serbian and Serbo-Croatian (dubbed sr-sh-nlp) developed in the scope of the AI PhD programme at the University of Belgrade contain resources prepared using the TEI Lex0 standard with 40,000 synonyms entries and more than 53,000 definition entries extracted from Serbo-Croatian Wiktionary data and data from the Systematic Dictionary (2,391 groups representing senses. Each group belongs to one of 18 fields (everyday life, psychology, ethics, mathematics, language, etc).[17]

Several multilingual lexical resources were built that include Serbian, such as *senti_lex* and *JULIELab/MEmoLon (Data)* aimed for sentiment analysis. Quite a number of multi- or bi-lingual terminology resources were developed, e. g. *ms_terms*, the Microsoft Terminology Collection, English-Serbian terms extracted from the English-Serbian INTERA corpus (see footnote 11), Eurovoc that covers 30 languages (with 9,000 entries having Serbian terms), and Agrovoc including almost 25,000 Serbian terms.

Systematic development of terminology resources has been done by the UB/JeRTeh researchers, and users can search them on the platform Termi (Kitanović et al., 2021).[18] These include more than 12,000 bilingual English/Serbian term pairs from various domains, and much more Serbian terms. A part of these resources (resources from the domains of geology, mining, power engineering) are registered as open for download. Terminology development by this group is an ongoing activity.

Two special purpose bilingual lexicons were developed by UB/JerTeh researchers, aimed to solve specific problems in sentiment analysis. One is the improved version of sr-HurtLex for abusive language detection in which a number of entries were corrected and more than 1,600 were newly added. For further development of this lexicon the AbCoSER corpus was built consisting of 6,436 tweets out of which 1,416 were annotated as containing the abusive speech, further classified in more fine grained categories (Jokić et al., 2021). The other is the sentiment lexicon for sentiment analysis that was obtained through the harmonisation, translation and adaptation of several multilingual sentiment lexicons (EmoLex, Bing, AFINN). The Serbian headwords in this lexicon (25,000) are supplied with POS and sentiment categories and aligned with used dictionaries. Both resources are still worked on and they will be published soon.

Proprietary resources were developed for similar purposes at the University of Niš: a hate-speech lexicon with 4,705 entries, collocations, phrases and idioms categorised into three groups: curses, insults and threats (Mladenović et al., 2020). It is accompanied by a cor-

[16] SrpMD4Tagging – Serbian Morphological Dictionaries for Tagging at ELG (https://www.european-language-grid.eu)
[17] NLP resources for Serbian and Serbo-Croatian *sr-sh-nlp* (https://github.com/putnich)
[18] Termi (https://termi.rgf.bg.ac.rs)

pus of almost 6,000 short comments on several portals, manually labelled as “hate/not hate speech/unknown”. In addition, the sentiment lexicon was built containing positive and negative words and phrases, and a dataset with tweets manually labeled as “ironic/not ironic”.

A rare bilingual lexicon that includes a language other than English is the Serbian/German lexicon consisting of 3,373 aligned concepts derived from the Serbian/German aligned literary corpus. This resource is open and its further development is frozen at this moment (Andonovski et al., 2019).

The Serbian WordNet, produced by researchers from UB is aligned to the Princeton WordNet 3.0, and with SentiWordNet. It has 22,571 synsets, and some domains are better populated than the others[19] Its last version is from 2018 (Stanković et al., 2018). A formal OWL2 domain ontology with 98 rhetorical figures accompanied by examples, classified into 4 rhetorical types was developed and made public (present status unknown).

Models and grammars

The Dict2Vec embedding model was adapted for Serbian in the scope of the AI PhD programme at the UB using Serbo-Croatian Wikipedia and Wiktionary synonym pairs to improve the training process. On average, using Dict2Vec increased word similarity scores between 55-70 percents, depending on the training configuration used.[20]

BERTić, available through HuggingFace is a transformer model pre-trained on 8 billion tokens of crawled text from the Croatian, Bosnian, Serbian and Montenegrin web domains, evaluated on POS tagging, NER, geo-location prediction and commonsense causal reasoning, showing improvements on all tasks over state-of-the-art models (see footnote 11) (Ljubešić and Lauc, 2021).

4.2 Language Technologies and Tools

Text analysis

Several taggers and/or lemmatisers for Serbian were developed, three of them by UB/JeRTeh researchers: TreeTagger was used for a long time as the optimal tagging approach, but the spaCy framework, which uses contemporary ML technology proved as a promising alternative, as well as the NLTK library POS-tagger, but with less satisfying results. Four tagging models, TreeTagger and spaCy trained on the same dataset with 2 different tagsets, UD and traditional Serbian POS tagset, as well as lemmatisers for these taggers supported by dictionaries, are published on the JeRTeh portal and on ELG (Stanković et al., 2020).[21] A complete suite for sentence segmentation, tagging, lemmatisation and other processing is also available on the web for public use.[22]

A suite of NLP tools for Serbian was developed and made public, it includes ParCoTrainSynt, a 101,000 token treebank, ParCoLex: a morphosyntactic lexicon with 6 million entries, two parsing models, a morphosyntactic tagging model, and a lemmatisation model.[23] Also, UDPipe performs segmentation, tokenisation, POS tagging, morphological analysis, lemmatisation and dependency parsing of raw Serbian texts (Miletic et al., 2019).

Text analysis of Serbian texts has been done for years in Unitex/Gramlab environment[24] using SrpMD. In this environment various local grammars were developed by UB/JerTeh researchers, e. g. for compound verb forms, nominal phrases etc. which were used to solve dif-

[19] It can be consulted at the site of the Bulgarian Academy of Sciences (http://dcl.bas.bg/bulnet)
[20] Dict2Vec (https://github.com/putnich)
[21] UB/JeRTeh taggers and lemmatisers (https://github.com/procesaur/srpski)
[22] UB/JeRTeh processing suite (https://github.com/petar-popovic-bg/Jerteh)
[23] Serbian-nlp-resources (https://github.com/aleksandra-miletic/serbian-nlp-resources)
[24] Unitex/Gramlab corpus processing suite (https://unitexgramlab.org)

ferent problems (NER (Šandrih Todorović et al., 2021), terminology extraction (Šandrih et al., 2020), definition modelling for dictionary entries (Stanković et al., 2021), de-identification of sensitive texts, etc.). Using the same environment, special-purpose grammars were built for the extraction of mining entities from domain texts, e. g. mining equipment, locations of the mine sites, equipment failure and failure entities.

Parsing of Serbian is possible online using UDPipe,[25] as a REST service, and is also implemented as an ELG compatible service. The CLARIN Knowledge Centre for South Slavic languages (CLASSLA) offers the pipeline for tokenisation and sentence splitting, POS tagging, lemmatisation, dependency parsing, NER for the processing of Serbian, among other languages.[26]

Several sentiment analysis systems were developed for Serbian that analyse movie reviews and news portals.

Speech processing

A substantial breakthrough in the area of speech processing was made by a group from the Faculty of Technical Sciences at the University of Novi Sad (Delić et al., 2019). They have commercialised their speech recognition and generation resources and tools by the AlphaNum company, a spin-off of the University of Novi Sad.[27] They offer a large variety of products and services: speech technologies (ASR and TTS), voice assistant, products for disabled, audiomemo recording, dictation, voice dial. Resources produced by these group are not made available, commercially or otherwise. Only few other resources exist, e. g. the spoken corpora *Serbian emotional speech database* (GEES), however, these are only available commercially.

Translation technologies

As for MT systems, besides some rudimentary attempts done in the scope of scientific research and products created by "big players" such as Google, there are MT models for German-Serbian and English-Serbian language pairs (HelsinkiNLP – OPUS-MT).[28]

Information Extraction and Information Retrieval

The UB/JeRTeH group developed several NER systems for Serbian. Besides the rule- and lexicon- based system *SrpNER* that tags fine-grained entities, which works in the Unitex/Gramlab environment (Krstev et al., 2014), several systems were developed using various ML methods and tools. One of them is SrpCNNER trained for seven classes of NEs.[29]

A web service was developed for query expansion, morphological and semantic, that was incorporated in several on-line applications, such as *Bibliša* digital library and the Geological Projects Library of the Ministry of Mining and Energy.

Some of the mentioned NLP applications and services, and more are available on the JeRTeh portal that is permanently updated (Šandrih Todorović et al., 2021),[30] e. g. BilTE for the bilingual terminology extraction (Šandrih et al., 2020), NER&Beyond[31] for the harmonisation of different annotation tagsets and Serbian NER models, feature extraction for good dictionary examples (Stanković et al., 2019). Several students projects are also published:

[25] UDPipe (https://lindat.mff.cuni.cz/services/udpipe/)
[26] CLASSLA (https://pypi.org/project/classla/)
[27] AlphaNum (https://www.alfanum.co.rs)
[28] See Footnote 11.
[29] SrpCNNER – Named Entity Recogniser for Serbian at ELG (https://www.european-language-grid.eu)
[30] JeRTeh portal (http://portal.jerteh.rs)
[31] NER&Beyonf (http://nerbeyond.jerteh.rs)

Baby WordNet: Python Wordnet Interface for English; Hate Speech Detection for Serbian based on developed models and lexicons; and Text Generator for Serbian.

4.3 Projects, Initiatives, Stakeholders

The document *Strategy for the Development of Artificial Intelligence in the Republic of Serbia for the period 2020-2025* was adopted by the Government in 2019 (StrategyAI, 2019). In this document, NLP is recognised as one of the important AI applications, and it is emphasised that priorities should also include fields that do not necessarily have an immediate economic effect, such as language. One of the goals set in this document is the foundation of the Institute for AI[32], which is in its initial stage of its work, with natural language understanding as one of its research areas.

However, there is still no LT-related funding in Serbia. The funding of national scientific and research projects has changed in the last two years, and they are now under the responsibility of the Science Fund of the Republic of Serbia and the Republic of Serbia Innovation Fund. In the few calls issued so far, just a few projects related to languages were found, although there were a number of applications: in the scope of the call, *Ideas* (dedicated to all sciences) two projects related to Serbian were adopted, but neither of them has LT as its primary goal. In the scope of the *AI* call one project related to speech processing and one remotely related to LT were chosen for funding, while in the scope of the *Promis* call dedicated to young researchers, the chosen projects are related neither to the Serbian language nor to LT for Serbian.

Researchers in Serbia, mostly affiliated with the state universities, primarily at the Universities of Belgrade, Novi Sad and Niš and research institutes (Institute of the Serbian Language of the Serbian Academy of Sciences and Arts), continue their research in LT for Serbian, either unfunded or funded by some short-term applied research projects (e. g. Wikipedia, government ministries, etc.). In the last five years these researchers were not involved in major international projects, except for ELG, but they participated in some bilateral projects (e. g. German-Serbian with the topic of hate-speech), and are very active in a number of COST actions: IC1207 “PARSEME: PARSing and Multi-word Expression” 2013-21017, CA16204 “Distant Reading for European Literary History” 2017-2022, CA16105 “NexusLinguarum” 2017-2021, CA18209 “European network for Web-centred linguistic data science” 2019-2023, CA19102 “Language In Human Machine Era (LIHTME)” 2020-2024.

Outside academia there are few LT providers. The largest provider is JeRTeh, whose resources, tools and services were described in Section 4.2. One can mention companies such as AlphaNum, providing commercial speech technologies and Lexicom, providing some LTs (mostly lexicons). When government institutions, as well as large national and international companies have a need for LT, they either try to find solutions offered by the researchers doing LT already, or more often they try to produce in-house solutions.

5 Cross-Language Comparison

The LT field[33] 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

[32] Institute for Artificial Intelligence (https://www.ivi.ac.rs)
[33] This section has been provided by the editors.

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[34] 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
  - 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 $\geq$3% and $<$10% of the ELG resources of the same type

[34] 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.

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[35]

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[36] 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 calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[37]

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

[35] 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.

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

[37] 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.

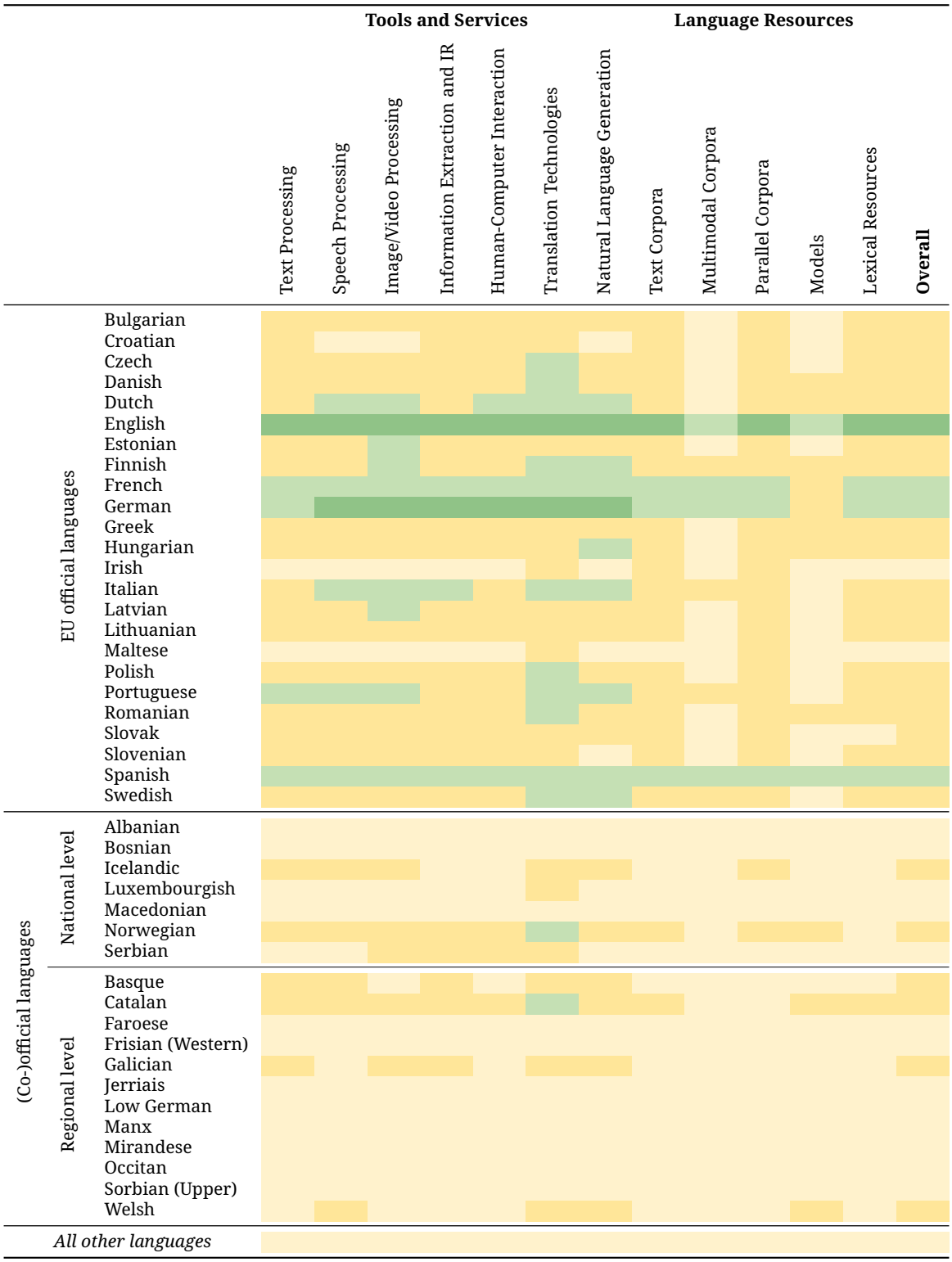

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

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,[38] 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.

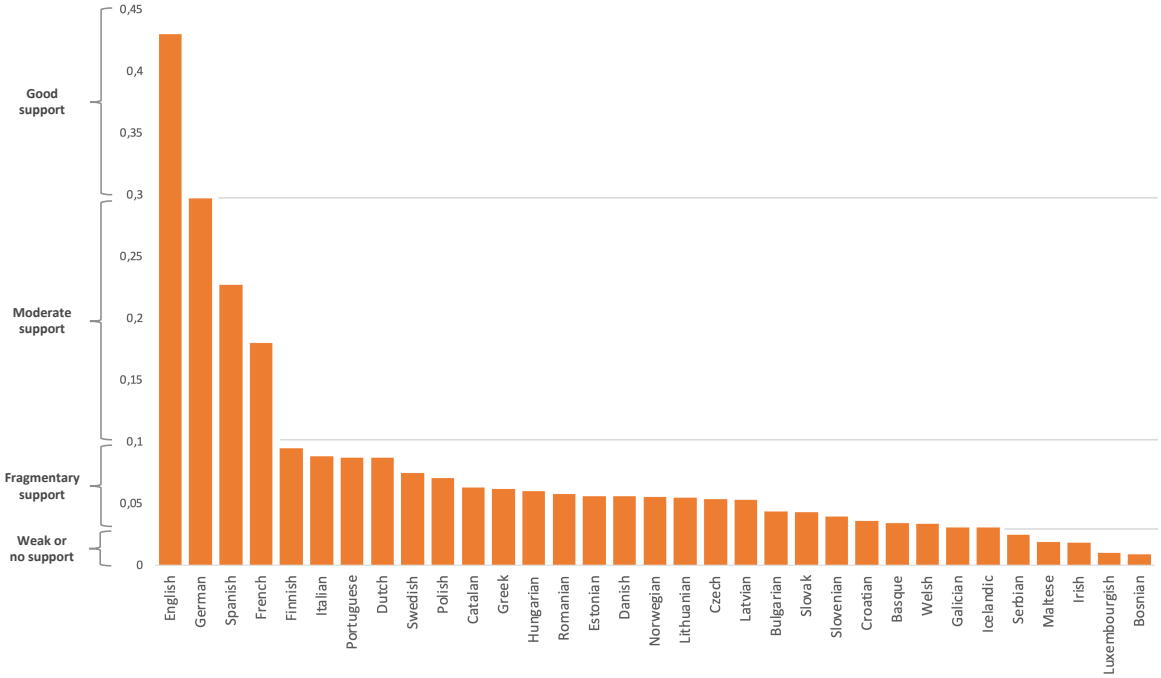

Figure 1: Overall state of technology support for selected European languages (2022)

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 im-

[38] 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.

balance, 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 direct 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

The overview presented in this deliverable shows that LT for the Serbian language is by no means neglected. There are research groups that have worked steadily for many years on producing language resources, tools and applications for Serbian, and new groups and individuals emerge in the field in recent years. As a result of those development some very valuable, comprehensive and robust resources and tools were developed. This is true especially for text corpora, parallel corpora and lexical resources. Speech technologies are well developed and many valuable applications were built upon them. Several POS taggers, lemmatisers, as well as NER systems were developed based on state-of-the-art methodologies. However, in many areas only insufficient or rudimentary resources and tools exist, such as multimodal corpora, grammars, language models, language generators.

This investigation enabled the comparison of LT for Serbian with English and a few other well supported languages. Although one can notice certain progress in LT for Serbian in the last ten years which resulted in the development of more and better resources and tools, there has been progress in the LT field as a whole that yielded even better resources and tools for English and a few other languages that were already in the lead. So today one can observe that despite the progress that has been achieved for Serbian the distance between Serbian and the well-supported languages has not been reduced. The research also revealed that the distance of languages close to Serbian spatially, historically and by a number of speakers, such as Bulgarian, Slovene and Croatian to moderate and good supported languages is smaller, though still large. The policies taken in corresponding countries to promote and to support LT for their languages can serve as a guide on how to improve LT for Serbian.

Despite valuable achievements Serbian is still a disadvantaged language that carries the danger that in a few years a Serbian speaker will be excluded from the digital sphere and s/he will not benefit from the upcoming AI/LT revolution. To prevent this, there is need for more funding, on both the national and international level. Although the importance of AI has been recognised in Serbia, and there has been funding dedicated to it, what is needed is funding dedicated to LT, whether in the scope of AI or not (e. g. theoretical research that would produce grammars for Serbian applicable to LT). At the international level, Serbian and other weakly supported languages would benefit from more knowledge transfer projects that would not aim at mirroring existing solutions for English but rather support production of adequate resources and tools for endangered languages.

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.

Jelena Andonovski, Branislava Šandrih, and Olivera Kitanović. Bilingual lexical extraction based on word alignment for improving corpus search. *The Electronic Library*, 2019.

Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. Fine-grained semantic textual similarity for serbian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

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

Vlado Delić, Zoran Perić, Milan Sečujski, Nikša Jakovljević, Jelena Nikolić, Dragiša Mišković, Nikola Simić, Siniša Suzić, and Tijana Delić. Speech technology progress based on new machine learning paradigm. *Computational intelligence and neuroscience*, 2019, 2019.

ICT2020. Употреба информационо-комуникационих технологија у Републици Србији, 2020. (The use of Information and Communication Technologies (ICT) in the Republic of Serbia in 2020), 2020. URL https://publikacije.stat.gov.rs/G2020/Pdf/G202016015.pdf.

Danka Jokić, Ranka Stanković, Cvetana Krstev, and Branislava Šandrih. A Twitter Corpus and Lexicon for Abusive Speech Detection in Serbian. In Dagmar et al. Gromann, editor, *3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *Open Access Series in Informatics (OASIcs)*, pages 13:1–13:17, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-199-3. doi: 10.4230/OASIcs.LDK.2021.13. URL https://drops.dagstuhl.de/opus/volltexte/2021/14549.

Olivera Kitanović, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić, Ivan Babić, and Ljiljana Kolonja. A data driven approach for raw material terminology. *Applied Sciences*, 11(7):2892, 2021.

Cvetana Krstev. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. University of Belgrade, Faculty of Philology, Belgrade, 2008.

Cvetana Krstev, Dusko Vitas, and Tomaz Erjavec. Morpho-syntactic descriptions in multext-east-the case of serbian. *Informatica (Slovenia)*, 28(4):431–436, 2004.

Cvetana Krstev, Ivan Obradović, Miloš Utvić, and Duško Vitas. A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, 24(2):473–489, 2014.

Nikola Ljubešić. Text collection for training the bertić transformer model bertić-data. 2021.

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 Davor Lauc. BERTić-The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, 2021.

Aleksandra Miletic, Cécile Fabre, and Dejan Stosic. De la constitution d'un corpus arboré à l'analyse syntaxique du serbe. *Traitement Automatique des Langues*, 2019.

Miljana Mladenović, Vladimir Momčilović, and Ivan Prskalo. Stl4nlp–web tool for manual semantic annotation of digital corpora. *The Strategic Directions of the Development and Improvement of Higher Education Quality: Challenges and Dilemmas*, page 200, 2020.

Ljubomir Popović. From standard Serbian through standard Serbo-Croatian to standard Serbian. In Celia Hawkesworth and Ranko Bugarski, editors, *Language in the former Yugoslav lands*, pages 25–40. Slavica Pub., 2004.

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.

Ranka Stanković, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. *Keyword-Based Search on Bilingual Digital Libraries*, pages 112–123. Springer International Publishing, Cham, 2017. ISBN 978-3-319-53640-8. doi: 10.1007/978-3-319-53640-8_10. URL http://dx.doi.org/10.1007/978-3-319-53640-8_10.

Ranka Stanković, Branislava Šandrih, Rada Stijović, Cvetana Krstev, Duško Vitas, and Aleksandra Marković. Sasa dictionary as the gold standard for good dictionary examples for serbian. *Electronic lexicography in the 21st century: Smart lexicography*, pages 248–269, 2019.

Ranka Stanković, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. Resource-based WordNet Augmentation and Enrichment. In Svetla Koeva, editor, *Proceedings of the Third International Conference Computational Linguistics in Bulgaria (CLIB 2018)*, pages 104–114. Institute for Bulgarian Language "Prof. Lyubomir Andreychin", Bulgarian Academy of Sciences, May 2018. ISBN 2367-5675 (on-line).

Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3947–3955, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.487.

Ranka Stanković, Cvetana Krstev, Rada Stijović, Mirjana Gočanin, and Mihailo Škorić. Towards automatic definition extraction for serbian. volume 2, page 695–704. Democritus University of Thrace, 2021.

StrategyAI. Strategy for the Development of Artificial Intelligence in the Republic of Serbia for the period 2020-2025 , 2019. URL https://www.media.srbija.gov.rs/medsrp/dokumenti/strategy_artificial_intelligence.pdf.

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.

Ustav. *Constitution of the Republic of Serbia*. 2006. URL http://www.ustavni.sud.rs/page/view/sr-Latn-CS/70-100028/ustav-republike-srbije.

Duško Vitas and Cvetana Krstev. Processing of corpora of Serbian using electronic dictionaries. *Prace Filologiczne*, 63:279–292, 2012.

Duško Vitas, Ljubomir Popović, Cvetana Krstev, Ivan Obradović, Gordana Pavlović-Lažetić, and Mladen Stanojević. *Српски језик у дигиталном добу – The Serbian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, 9 2012. URL http://www.meta-net.eu/whitepapers/volumes/serbian. Georg Rehm and Hans Uszkoreit (series editors).

Branislava Šandrih, Cvetana Krstev, and Ranka Stanković. Two approaches to compilation of bilingual multi-word terminology lists from lexical resources. *Natural Language Engineering*, 26(4):455–479, 2020. doi: 10.1017/S1351324919000615.

Branislava Šandrih Todorović, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. Serbian ner& beyond: The archaic and the modern intertwined. In Galia et al. Angelova, editor, *Deep Learning Natural Language Processing Methods and Applications – Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1252–1260. INCOMA Ltd., September 2021. ISBN 978-954-452-072-4. URL https://doi.org/10.26615/978-954-452-072-4_141.