EUROPEAN LANGUAGE EQUALITY

D1.5

Report on the Bulgarian Language

Authors	Svetla Koeva, Valentina Stefanova
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	D1.5
Deliverable title	Report on the Bulgarian Language
Type	Report
Number of pages	27
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 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	Svetla Koeva, Valentina Stefanova (Sections 2, 4 and 6)
Reviewers	Maria Giagkou, Teresa Lynn
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
CALL	Computer-assisted language learning
CEF	Connecting Europe Facility
CEF AT	Connecting Europe Facility, Automated Translation
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CPU	Central Processing Units
CURLICAT	Curated Multilingual Language Resources for CEF AT
DLE	Digital Language Equality
EC	European Commission
EFNIL	European Federation of National Institutes for Language
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
EU	European Union
GPU	Graphics Processing Unit
HPC	High Performance Computing
ICT	Information and Communications Technology
IT	Information Technology
LIMA	Libre Multilingual Analyzer
LR	Language Resource/Resources
LT	Language Technology/Technologies
MARCELL	Multilingual Resources for CEF.AT in the legal domain
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NLP	Natural Language Processing
NLG	Natural Language Generation
NLU	Natural Language Understanding
PARSEME	PARSing and Multi-word Expressions
POS	Part-of-Speech
SME	Small and Medium-sized Enterprise

SR	Speaker Recognition
SRA	Strategic Research Agenda
STT	Speech-to-Text
TTS	Text-to-Speech
UD	Universal Dependencies

Abstract

This document reports on the current status of technology support for the Bulgarian language and highlights the identified gaps that should be overcome by further development of research and technology.

The Bulgarian language, the official language of the Republic of Bulgaria, is spoken by over eight million native speakers, mainly in Bulgaria, but also in Europe and North America. The mixture of specific language characteristics (the rather flexible word order, the lack of morphological distinction for nominal cases and subject omission) makes the Bulgarian language a real challenge for natural language processing.

Language Technology (LT) is the multidisciplinary scientific and technological field that is concerned with studying and developing systems capable of processing, analysing, producing and understanding human languages, irrespective of their being written, spoken or embodied. LT is trying to provide solutions for the following main application areas: Text Analysis; Speech Processing; Machine Translation; Information Extraction and Information Retrieval; Natural Language Generation; Human-Computer Interaction.

The study of the available language technology for Bulgarian is based on the overall collection recorded on the European Language Grid (ELG) in February 2022. The observations are briefly as follows. Freely available monolingual and parallel corpora in Bulgarian appeared only recently. Large multilingual corpora which include Bulgarian data are usually created automatically from the Internet (often from Wikipedia). Annotated Bulgarian corpora containing manually validated or manually assigned linguistic information are smaller in number and in volume and most of them are monolingual. There are not enough speech and multimodal corpora and most of them are multilingual, thus not specially designed for Bulgarian.

Text analysis still dominates the field of Bulgarian language technology, and multimodal input data, such as simultaneous text, images, audio and video, are rarely processed, although forecasts indicate that video content will soon dominate on the internet. There are also applications for automatically translating language, although these still fail to produce linguistically and idiomatically correct translations, especially when Bulgarian is the target language.

Many commonly used and necessary technologies are still not available for Bulgarian (human-computer interaction, multimodal processing, etc.) and for others, even if some advance in technologies is recorded, there are no certain available applications (summarisation, question answering, etc.). Many technologies are advanced abroad and Bulgarian is part of some multilingual systems for machine translation, speech analysis and recognition.

The general conclusion is that there is still a yawning technological gap between English and Bulgarian and even between German, French, Italian, Spanish and Bulgarian. A comparison of international technology and the one for Bulgarian shows that results for the automatic analysis of English and of some other European languages are far better than those for Bulgarian.

The LT field as a whole has significantly progressed in the last ten years. 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.

Резюме

Изследването представя съвременното състояние на езиковите технологии за български език, като се обръща специално внимание на областите, в които езиковите

ресурси и технологии за български са слабо развити. Българският език е официалният език на Република България и се говори от над 8 милиона души основно в България, но също така в Европа и в Северна Америка. Българският е първият славянски език, който разполага със собствена писмена система, датираща от 9. век. На 1 януари 2007 г., когато България е приета за пълноправен член на Европейския съюз, българската азбука става третата официална азбука на Европейския съюз след латинската и гръцката. Смесването на специфични езикови характеристики (относително свободен словоред, липса на падежни окончания и изпускане на подлога) прави българския език истинско предизвикателство за компютърната обработка на езика.

Езиковите технологии обхващат широка интердисциплинарна научна област, която се занимава с изучаването и разработването на системи, способни да обработват, анализират, възпроизвеждат и „разбират“ човешките езици, независимо дали са в писмена, или в устна форма. Езиковите технологии предоставят решения за следните основни области на приложение: автоматичен анализ на текста, обработка на реч, машинен превод, автоматично извличане на информация, генериране на естествен език, взаимодействие между човек и компютър.

Изследването на съществуващите към момента езикови технологии за български език се базира на колекцията от ресурси, програми за обработка на езика и услуги, събрани и разпространени от проекта „Европейска езикова мрежа“ (ELG) до февруари 2022 г. Наблюденията накратко са следните. Свободнодостъпни едноезикови, паралелни и многоезикови корпуси на български език се появяват сравнително неотдавна. Големите многоезикови корпуси, които включват български данни, в повечето случаи са създадени автоматично от текстове, които са достъпни в интернет (често от Уикипедия). Анотираните български корпуси, съдържащи ръчно проверена езикова информация, са по-малко на брой и с по-малък обем, като повечето са едноезикови. Няма достатъчно корпуси на българска реч или корпуси с многомодално съдържание (обединяващо текст, реч и изображения), а голяма част от достъпните за използване корпуси са многоезикови и не са специално създадени за български език.

В областите на приложение на езиковите технологии за български доминира анализът на текста, а обработката на многомодални данни се предлага рядко. Съществуват и редица приложения за автоматичен превод от и на български език, но все още автоматичният превод не се характеризира с отлично качество особено когато текстовете включват преносни и идиоматични изрази, а българският е целевият език.

Често използвани и нужни на хората езикови технологии не са разработени за български език (например за взаимодействие между човек и компютър, за едновременна обработка на текст, реч, изображения или видео), а за други, макар че има известен напредък на технологично равнище, няма разработени приложения за широко използване (например за автоматично резюмиране на съдържанието на документи, за автоматично отговаряне на въпроси и др.). Изследванията по отношение на много технологии са напреднали в чужбина и българският е част от някои многоезикови системи за машинен превод, анализ и разпознаване на реч.

Изводът, който може да се направи, е, че все още се наблюдава съществена разлика между езиковите технологии не само за английски и български език, но и за немски, френски, испански и български език. Съпоставката на съществуващите езикови приложения показва, че автоматичната обработка за английски език (а и за някои други европейски езици) е с много по-добро качество в сравнение с тази за български език.

Езиковите технологии като цяло бележат значителен напредък през последните десет години. Разликата обаче между езиците с развити езикови технологии и езиците със слаборазвити езикови технологии се запазва и през 2022 година. Тази разлика трябва да бъде преодоляна (или на първо време значително намалена), за да се осигури равнопоставеност на европейските езици по отношение на възможностите за компютърна обработка, основана на езикови технологии.

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.

This document reports on the current status of the technology support for the Bulgarian language and highlights the identified gaps that should be filled up by further development of research and technology.

2 The Bulgarian Language in the Digital Age

2.1 General Facts

Bulgarian is the official language of the Republic of Bulgaria. It is spoken by over eight million native speakers, mainly in Bulgaria, but also in Australia, Canada, Germany, North Macedonia, Spain, Turkey (Europe), Ukraine, the United Kingdom, and the USA. It is also spoken in the Austria, Czech Republic, France, Greece, Italy, Israel, Moldova, Romania, the Russian Federation (Europe), Serbia, and Slovakia. A recent report of the World Bank[3] states that approximately 1.7 million Bulgarians lived abroad in 2020. According to a preliminary assessment made by the National Statistical Institute for the 2021 census, the population in Bulgarian is about 6,500 000[4] and the proportion for Bulgarian native speakers remains relatively constant – about 85% of the population (Koeva, 2018).

The official alphabet is Cyrillic. Bulgarian is the first Slavic language with its own writing system, which dates from the 9th century. The first Old Bulgarian alphabet, the Glagolitic script, was created by St. Cyril the Philosopher and most scholars accept the year of creation of the Glagolitic letters to be 855, as stated in the work “On the Letters” by Chernorizets Hrabar. At the end of the 9th and the beginning of the 10th century the second Bulgarian alphabet, the Cyrillic script, was created at the Bulgarian literary schools and introduced in the First Bulgarian Kingdom (Мичева et al., 2021). The Bulgarian alphabet has been the third official script in the European Union since 2007, when Bulgaria became a member.

There are many regional Bulgarian dialects, colourful varieties of the Bulgarian language spoken both inside and outside the borders of the country.

[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
[3] https://openknowledge.worldbank.org/bitstream/handle/10986/36800/Migration-in-Bulgaria-Current-Challenges-and-Opportunities.pdf
[4] https://census2021.bg/новини/по-предварителна-оценка-населението/

Particularities of the Bulgarian language

Bulgarian belongs to the family of South Slavic languages and forms part of the Balkan linguistic union (Balkan Sprachbund). Consequently, Bulgarian displays similarities to both language groups. As a Slavic language, Bulgarian possesses a rich inflectional and derivational morphology, verb aspect pairs, etc. However, due to the mutual influence of Balkan languages, Bulgarian has lost noun cases (except vocative) and has completely lost the infinitive form.

Bulgarian exhibits a number of specific characteristics that contribute to the richness of the language but can also be a challenge for the computational processing of natural language. For example, it has a relatively free word order, while the order of adjectives as pre-positive modifiers of nouns or of adverbs as pre-positive modifiers of adjectives and other adverbs is fixed, the order of the subject and object of the verb is relatively loose. Unlike other languages which show a relatively loose word order (the other Slavonic languages, for example), Bulgarian does not possess nominal case inflection to indicate the syntactic relations between words. Yet another characteristic feature of Bulgarian which poses difficulty for syntactic parsing is the free omission of the subject which, when combined with the possibility of shifting the positions of the subject and object, makes the task even harder. The rather flexible word order, combined with the lack of morphological distinction for nominal cases and subject omission is a real challenge for the natural language processing of Bulgarian.

Official language protection in Bulgaria

The Bulgarian constitution states that Bulgarian is the official language in the Republic of Bulgaria. Constituted by a decree of the Council of Ministers, the Institute for Bulgarian Language of the Bulgarian Academy of Sciences is the official institution which monitors changes in the Bulgarian language, determines literary norms and reflects these changes in both orthography and grammar. Its primary tasks include research in Bulgarian linguistics, general, theoretical, applied and computational linguistics, as well as the preparation of a comprehensive dictionary of the Bulgarian language. Other research projects investigate Bulgarian dialects within and outside Bulgaria, including issues of language policy within the framework of European integration.

The Radio and Television Act states that the Bulgarian National Radio shall set aside for the creation and performance of Bulgarian music and radio drama not less than 5% of the subsidy of the state budget; while the Bulgarian National Television shall set aside no less than 10% of the same subsidy for Bulgarian television film production. The wide usage of Language Technology can make an important contribution in this area by offering media, internet and mobile communications sophisticated language services.

Language in education

Since the 19th century, Bulgarian language and literature have had a very important role in institutionalised education. According to the legislation in Bulgaria, all education and teaching provided as part of the current state curriculum, from pre-school through to university level, must be in Bulgarian. The study of Bulgarian is compulsory for elementary and secondary school. Increasing the volume of Bulgarian language teaching in schools is one possible step towards providing students with the language skills required for active participation in society.

3,886,740 or 49% of people spoke foreign languages in 2020 according to the National Statistical Institute, among which 1,922,190 speaking at least one foreign language, and the rest more. Language Technology can make an important contribution here by enhancing

Computer-Assisted Language Learning (CALL) systems, with which students learn language through play.

International aspects

Education of Bulgarians Abroad and School Network Directorate at the Ministry of Education and Science is responsible for the state policy related to the education and training of Bulgarians living outside the Republic of Bulgaria. There are 400 schools in 40 countries on six continents for Bulgarian children abroad (as of 2021). Despite the co-financing and the methodological support by the state, they exist most of all because of the efforts, the resources and the enthusiasm of fellow Bulgarians. For many years the Ministry of Education and Science has organised for Bulgarian university lecturers to work as lecturers in the Bulgarian language, literature and culture in a number of foreign universities where Bulgarian is studied. The Ministry of Culture established the National Culture Fund, which holds regular competitions for the translation of Bulgarian literature into foreign languages. Overseas, outside certain narrow specialised circles and Bulgarian communities abroad, Bulgarian is an unknown and exotic language. There is still no Bulgarian cultural institute established to promote Bulgarian culture, language and history in other countries, such as the Goethe Institute, the Cervantes Institute, etc. Language technologies could help European citizens to get acquainted with the Bulgarian history and culture, as well as Bulgarians to establish and expand their business, cultural and scientific relations abroad.

Bulgarian has acquired the status of an official administrative language of the European Union on the same basis as English, German and French, since the European Union is based on solidarity and equality amongst its members. The fact of Bulgaria's membership in the European Union together with the idea of unity and diversity, globalisation while preserving national identity, provides a real opportunity for the egalitarian use of Bulgarian together with the major European languages.

2.2 Bulgarian in the Digital Sphere

Bulgarian internet users in 2020 increased by 31% in comparison to 2007 and already 46% of the total population use the internet. According to data provided by the National Statistical Institute, around the end of 2020 78.9% of the households in Bulgaria have access to the internet and the people using the internet on a regular basis (every day or at least once per week) amount to 69.2%. 30.9% of the population in Bulgaria have bought goods and services online for personal use in the last 12 months. Diverse reasons for not shopping online have been pointed out; among others the lack of knowledge and skills, including lack of knowledge of a foreign language, as indicated by 5.4% of people.

According to W3Techs,[5] Bulgarian figures 0.1% as content language on the web (as in November 2021). Among the popular web search engines, social media and applications that use Bulgarian are: Google search, Microsoft Bing, Facebook, Wikipedia, Whatsapp, Wordpress, Booking.com; Blogger.com. Among the 50 top websites in Bulgarian, 32 websites use the .bg country code as top-level domain, the most popular of which are news portals. Bulgarian Wikipedia, as an important source of data for natural language processing, contains 278,943 articles (as in February 2022), a considerably smaller size than the biggest Wikipedias – English, German and French – yet still in the 39th position among the Wikipedias in other languages.

[5] https://w3techs.com

3 What is Language Technology?

Natural language[6] 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 communities work closely together, combining methods and approaches inspired by both, together making up *language-centric Artificial Intelligence* (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.

[6] 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).

- **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 (TTS) 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.[7]

4 Language Technology for Bulgarian

The European Language Grid is now a central point for accessing Bulgarian language resources, tools and services. Much of these were sourced from databases like META-SHARE, ELRC-SHARE, ELRA Catalogue of Language Resources and SHARE, CLARIN, GitHub, Hugging Face, etc. The summary here is based on this overall collection recorded on the ELG. All data presented in the study are from February 2022.

[7] 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).

4.1 Language Data

Monolingual, bi- and multilingual text corpora

Technological developments in recent years and increase in CPU power enables the processing of huge language data, including large corpora, and allows the application of complex models and algorithms, which, undoubtedly, will lead to significant progress in the area. The Bulgarian National Corpus[8] is a large monolingual corpus for Bulgarian (comprising more than four billion tokens as of February 2022), enhanced with a detailed metadata description and linguistic annotation and designed to provide comprehensive information about the language.

The reference corpora are usually supplied with a corpus query system for online references; the same goes for the Bulgarian National Corpus with parts of it, with clear copyright, available for download.[9] There are also specialised monolingual corpora representing different domains: news, law, fiction, politics, business, competition, etc.

As of February 2022 we can identify 130 parallel text corpora, the majority of which are available for download from repositories such as ELG, ELRC-SHARE, and CLARIN. The bilingual corpora (87) mostly contain Bulgarian and English (52 corpora) or European language pairs, e. g. Bulgarian – Modern Greek, Bulgarian – German, Bulgarian – French, Bulgarian – Italian, Bulgarian – Spanish and so on, but also language pairs with some non-European languages such as Norwegian Bokmål, Persian, Ancient Greek (to 1453) and Latin.

Bulgarian is present in 159 multilingual corpora, and some of the multilingual corpora are sentence-aligned, which allows for easy cross-lingual research. Large multilingual corpora are usually created automatically from the Internet (often from Wikipedia) and are initiated by large companies or projects. There are also some comparable multilingual corpora. For example, MARCELL is a comparable corpus of national legislation documents for seven languages (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, and Slovenian). Each monolingual corpus is pre-processed (tokenised and morphologically tagged), classified and enriched with specialised terminology identified (Váradi et al., 2020). The Bulgarian MARCELL corpus consists of 29,648 legislative documents (at the end of March 2021), with a total of 61,226,208 tokens.

Annotated corpora containing manually validated or manually assigned linguistic information are smaller in number and in volume. Most of them are monolingual (24) and are designed for different purposes: Part-of-Speech tagging and Lemmatisation (for example, the Bulgarian Part-of-Speech Corpus),[10] Syntactic Parsing (for example, the Dependency part of BulTreeBank),[11] Word Sense Disambiguation (for example, the Bulgarian Sense annotated Corpus),[12] Named Entity Recognition (for example Bulgarian dataset for Named Entity Recognition).[13] Among the multilingual annotated corpora where Bulgarian is present, there are two large collections: Universal Dependencies treebank v2.8.1,[14] and annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.1),[15] both freely available for download. A collection of datasets for Bulgarian is available, some of the resources are automatically obtained, others are supplied with extensive annotation, and all are directed to a wide range of NLP applications.[16]

We can identify 21 spoken corpora for Bulgarian, seven of which contain both the tran-

[8] http://dcl.bas.bg/bulnc/
[9] https://dcl.bas.bg/bulnc/en/dostap/izteglyane/
[10] http://dcl.bas.bg/poscor/en/
[11] http://bultreebank.org/en/resources/short-description-dependency-part-bultreebank-bultreebank-dp/
[12] http://dcl.bas.bg/semcor/
[13] https://github.com/usmiva/bg-ner
[14] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3687
[15] https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2842
[16] https://huggingface.co/datasets?languages=languages:bg&sort=downloads

scriptions of spoken or spontaneous speech and the associated recordings, and nine only the transcriptions. Most of the spoken corpora for Bulgarian are multilingual.

There are very little examples of multimodal corpora, such is the TED-talks dataset, which contains filtered YouTube videos and speech recordings accompanied with transcriptions and subtitling.[17] Recently, a new multimodal dataset was developed containing over 20,000 images with object segmentation supplemented with textual data in Bulgarian and 25 more languages (Koeva, 2021).

In recent years, the large-scale mono- and cross-lingual parallel corpora have become the necessary prerequisites for current statistical and neural machine translation systems. The acquiring of high-quality monolingual and parallel corpora that are large enough to train neural models still remains an urgent task for small languages as Bulgarian and for domain-oriented or multimodal translation tasks.

Lexical/conceptual resources

Bulgarian is relatively well-resourced when it comes to electronic dictionaries, thesauri, and glossaries. Most dictionaries have been developed at the Institute for Bulgarian Language, however due to copyright restrictions, most of them only offer single user queries or data access for research purposes only, for example, the Dictionary of the Bulgarian Language (113,000 entries),[18] the Dictionary portal *LexIt* (12 dictionaries)[19] and some others. Some large multilingual knowledge bases, such as BabelNet,[20] or YAGO,[21] are fully available and have also support for Bulgarian. The information which they include is extracted from Wikipedia, WordNet,[22] and GeoNames,[23] and some other resources. The Bulgarian WordNet shares features of explanatory and specialised dictionaries and of an ontology: it includes 121,282 synonymous sets linking words and multiword expressions, which are classified according to a detailed system of stylistic and grammatical labels and semantic classes; the concepts represented by synonymous sets are linked with an extensive system of semantic, morpho-semantic, and extra-linguistic relations.[24] Parts of the Bulgarian WordNet are available for download, extended with semantic classes, new semantic relations and semantic frames.[25]

Models and grammars

There is a formal grammar of Bulgarian published in 2007 (Осенова and Симов, 2007). In addition, a finite-state transducers toolbox for Bulgarian accompanies the Bulgarian Grammatical Dictionary to allow for the generation of inflected forms of nouns, adjectives, verbs and prepositions.

Language models are important for developing natural language processing applications. However, training complicated NLP language models is a time-consuming process and there are not many available models for Bulgarian. A result of the transfer from Multilingual BERT model to several monolingual models, including Bulgarian with fine-tuning performed on Wikipedia data, is the Slavic BERT model (Arkhipov et al., 2019). Bulgarian is also among the languages with pre-trained word vectors, trained on Common Crawl and Wikipedia using

[17] https://huggingface.co/datasets/ted_talks_iwslt
[18] https://ibl.bas.bg/rbe/
[19] https://ibl.bas.bg/dictionary_portal/lang/en/
[20] https://babelnet.org
[21] https://yago-knowledge.org
[22] https://wordnet.princeton.edu
[23] https://www.geonames.org
[24] https://dcl.bas.bg/bulnet/
[25] https://dcl.bas.bg/en/semantichni-mrezhi/

fastText.[26] ELMo Representations models offer both complex characteristics of word use (e. g. syntax and semantics) and how these uses vary across linguistic contexts (i. e. to model polysemy) (Peters et al., 2018). Multilingual masked language models like RoBERTa base (Liu et al., 2019) and XLM-R (Conneau et al., 2020) address the cross-lingual understanding tasks by jointly pre-training large Transformer models on many languages, including Bulgarian.

Training of even larger models for Bulgarian and the development of sample-efficient pre-training settings become even more important with the increasing of the amount of available data and computing power. Such model representations will help to significantly improve any future efforts made towards tackling NLP problems for Bulgarian such as question answering, textual entailment, semantic role labeling, co-reference resolution, named entity extraction, and sentiment analysis.

4.2 Language Technologies and Tools

The ELG catalogue lists a number of multilingual tools and services that include Bulgarian as a supported language (e. g. for machine translation: Opus MT, SYSTRAN Translate, Tilde Machine Translation, or for linguistic analysis: LIMA – Libre Multilingual Analyzer, NLP-Cube; for spell-checking: GNU Aspel, etc.). Bulgarian is lacking in general in freely available robust speech and language tools designed specifically for Bulgarian.

Text Analysis

There are several NLP libraries providing sets of linguistic annotations (tokenisation, sentence splitting, paragraph boundary detection, spell checking, lemmatisation, named entity recognition, chunking, dependency parsing, sentiment analysis, etc.). Some of the libraries are Java-based such as OpenNLP and Stanford NLP, others are Python-based such as spaCy, and some support multiple programming languages, i. e. Spark NLP (Zaharia et al., 2016). A number of libraries provide deep learning techniques and knowledge graphs, and report good level of accuracy and speed (i. e. Spark NLP); however, they do not include pre-trained models for Bulgarian. Recently, two NLP pipelines (including a tokeniser, a sentence splitter, a tagger, a lemmatiser and a dependency parser) have become available: UDpipe (Straka and Straková, 2017) and NLP-Cube (Boroș et al., 2018). Both pipelines are trained for languages with the UD Treebanks, including Bulgarian. Another NLP pipeline for processing Bulgarian texts integrates a sentence splitter, a tokeniser, a part-of-speech tagger, a lemmatiser, a UD parser, a named entity recogniser, a noun phrase parser, and a term annotator (Koeva et al., 2020).

Text analysis still dominates the field of Bulgarian Language Technology (Hristova, 2021), and multimodal input data are rarely processed such as simultaneous text, images, audio and video, although forecasts indicate that video content will soon dominate on the internet. Moreover, the analysis of multimedia content based on only one medium is not reliable.

Speech Processing

While under development for several decades, the quality of speech technology is not yet satisfactory for a low-resourced language such as Bulgarian. Bulgarian speech corpora have been sporadically developed for specific research purposes, but they are usually smaller in size and heterogeneous in terms of formats and annotation: BGSpeech,[27] a corpus of transcribed conversational speech, and BulPhonC,[28] a corpus of declarative and interrogative

[26] https://fasttext.cc/docs/en/crawl-vectors.html
[27] http://bgspeech.net
[28] http://lml.bas.bg/BulPhonC/

sentences read by 147 different announcers, are some examples. SpeechLab 2.0 is a system for speech synthesis, specifically designed for Bulgarian. Several other systems integrate speech technologies for Bulgarian. The system eSpeak is an open source TTS system that can be further developed and improved, trained on large resources and easily integrated into various applications. Speech synthesis is also a functional task in the KNFB reader, a multilingual software for visually impaired people, that turns printed text into speech. The SkyCode TTS synthesiser is available for online use, as well as for download as an Android application. TTS for many languages with support for Bulgarian are also available, e. g. Speech Services by Google, Microsoft Speech Platform SDK, Vocaliser TTS, Acapela TTS, etc.

Although there are algorithms for speech recognition specifically designed for Bulgarian (Mitankin et al., 2009), there is still no publicly available Speech-to-Text (STT) system for Bulgarian. This leaves the niche of Speech Recognition for Bulgarian unoccupied, especially with a view to developing a system that can be integrated into different applications and platforms focused on social services in various areas of social life.

Within the project European Live Translator,[29] an automatic subtitling system of live meetings and conference presentations was developed providing spoken language translation (interpreting) from English to Bulgarian. Authôt Speech Recognition and Transcription[30] and Beey[31] are other proprietary services for automatic transcription of audio files based on Speech Recognition.

In general, there has been a development in the field of speech technologies for Bulgarian, but there are still no accessible and reliable STT systems, especially working in real time.

Translation Technologies

Excluding the automatic translation offered by multinational enterprise giants such as Google and Microsoft, there are other available machine translation systems from and into Bulgarian with free but limited access: Prompsit Translator, a free online service that uses a rule-based translator for related languages; SYSTRAN Translate, a free online translation service for text snippets, based on neural machine translation technology; the TraMOOC platform, a free translation service, that uses neural machine translation technology; DeepL Translator, based on neural networks and able to capture nuances and reproduce them in translation, etc. Custom neural networks machine translation engines from Bulgarian to English and the reverse were developed under the project CEF Automated Translation for the EU Council Presidency[32] funded by the Connecting Europe Facility of the EU Commission, the Telecommunications Programme, 2016-EU-IA-0121 I. The CEF eTranslation builds on the European Commission's earlier machine translation service, MT@EC, and applies neural machine translation in the 24 official EU languages. The CEF eTranslation is accessible (for public administration and SMEs) through API, a web user interface and the CEF Social Media Translator service.

Assessment of the quality of existing machine translation services, the number of language pairs, and the coverage of thematic domains still determines the machine translation technologies for the Bulgarian language as underdeveloped.

Information Extraction and Information Retrieval

Recently, there has been serious advances in research based on information extraction for Bulgarian: event extraction (Tanev and Steinberger, 2013); sentiment analysis (Kapukaranov

[29] https://elitr.eu
[30] https://www.authot.com
[31] https://www.beey.io
[32] https://www.tilde.com/products-and-services/machine-translation/de-presidency?lang=en

and Nakov, 2015); fake news detection (Dinkov et al., 2019; Karadzhov et al., 2018); result prediction (Velichkov et al., 2019); fact-checking (Atanasova et al., 2019). Some companies offer analytics, sentiment analysis, event extraction,[33] business intelligence,[34] and media monitoring.[35] For example, LIMA: Libre Multilingual Analyzer[36] is a multilingual linguistic analyzer and information extraction system.

Human-Computer Interaction

Popular systems such as Wolfram Alpha,[37] IBM Watson, as well as specialised systems such as EAGLi in Medicine[38] and others exist. Personal assistants such as Alexa or services such as OK Google or Siri do not support Bulgarian.

4.3 Projects, Initiatives, Stakeholders

The resources, tools and services for Bulgarian are distributed mainly by European repositories such as META-SHARE, ELRC-SHARE, ELG platform, CLARIN inventory and ELRA catalogue. Some of the resources which are available are developed at research centres (there are two main research centres in Bulgaria dedicated to LT). Some of the datasets are provided by the public administration, mainly thanks to the awareness raising activities of the ELRC project about the value of language data held by the public sector. There is only one national infrastructure dedicated to language resources, CLaDA-BG,[39] and it is focused on the storage and distribution of resources related to the linguistic, cultural and historical heritage. There are not many Bulgarian-language specific industry players in the speech and Language Technology space. Moreover, many companies in Bulgaria refuse to develop automatic translation and speech technologies because they could not compete with the international corporate giants in the market.

Many of the achievements and advancements in the development of language data and tools for Bulgarian have been the result of small, short-term funded and self-funded projects or PhD theses (e. g. spell-checker, grammar-checker, morphological databases, Part-of-Speech taggers, treebank, etc.). Despite the lack of dedicated funding and infrastructure for Bulgarian LT, there have been a number of notable projects that have helped shape the LT landscape into what it is today. Three CEF-funded projects have proven to be most impactful in the area of machine translation: CEF Automated Translation for the EU Council Presidency, MARCELL: Multilingual Resources for CEF.AT in the legal domain[40] and CURLICAT: Curated Multilingual Language Resources for CEF AT.[41] The European Language Resource Coordination (ELRC) has provided the opportunity for awareness-raising through outreach workshops (2016, 2017).[42] Overall, there is a good level of international cooperation and maintaining partnerships with leading research centers in the EU, although the funding is insufficient for using the full capacity of international cooperation.

Simultaneously, a growth of the high-tech Information Technology (IT) sector due to competitive labor prices is evidenced. The key facts on human resources availability are relatively positive: 220 high-schools with Information and Communication Technology (ICT)

[33] https://identrics.net/solutions/
[34] https://www.tetracom.com/ilib/tetracom/products/graze
[35] https://eventregistry.org
[36] https://github.com/aymara/lima/
[37] http://www.wolframalpha.com
[38] https://eagl.unige.ch/EAGLi/
[39] https://clada-bg.eu
[40] https://marcell-project.eu
[41] https://curlicat.eu
[42] https://lr-coordination.eu/pworkshops

focused curricula backed up with a solid language preparation and over 15 universities offering majors in ICT.[43] In 2017, there were more than 7,600 Information and Communication Technologies, and Engineering university graduates in Bulgaria. Between 200 and 250 of these graduates are estimated to be joining the AI field annually, according to the Data Science Society.[44] There are 930 IT companies operating in Bulgaria, among which 558 companies have offices in Bulgaria.[45] Around 50 companies are working within the AI domain and two-thirds of them are startups.[46] While the actors at research centres are well known, the dynamics in the development of the IT sector in Bulgaria is difficult to monitor. In a first study on the deployment of AI technologies across the EU,[47] the Bulgarian companies pointed out the lack of professionals in the field of big data management, machine learning and modelling. The same survey highlights that among the external barriers to the AI in Bulgaria the biggest are the lack of public and external funding and the need for new laws and regulations.

The ELG data shows 204 institutions and companies which have been developing language resources for Bulgarian and 49 of them are situated in Bulgaria.[48] A number of Bulgarian high-tech companies are very successful in the field of AI, for example Ontotext,[49] operating in the field of semantic technologies with its product GraphDB for managing knowledge graphs. However, the research centres and the IT companies work in isolation and most of the developed resources and tools are not reused or reintegrated, the links of scientific organisations with the business sector are weak and the mechanisms for knowledge transfer are insufficiently effective.

The Bulgarian AI strategy offers a policy vision for the development and use of AI in Bulgaria for the period 2020-2030 and identifies main areas of impact such as infrastructure and data availability, research and innovation capacity, knowledge and skills, and building trust in society (among them are the development of AI applications for educational purposes and systems for processing and communication in natural language). The creation of a Bulgarian AI Research Centre of Excellence is proposed to focus on the fields of neural networks, machine learning, natural language processing, semantic technologies, and robotics. There are also some funding opportunities for research and development in Bulgaria. The National Scientific Fund supports LT projects alongside all other disciplines. However, there is no dedicated funding for Bulgarian LT yet.

5 Cross-Language Comparison

The LT field[50] 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).

[43] https://www.investbg.government.bg/bg/sectors/human-resources-18.html
[44] https://investsofia.com/wp-content/uploads/2019/06/Artificial_intelligence_ecosystem_in_Bulgaria_2019-SeeNews_and_Vangavis.pdf
[45] https://dev.bg/company/
[46] https://www.trendingtopics.eu/ai-startups-corporates-communities-in-bulgaria/
[47] https://data.europa.eu/doi/10.2759/759368
[48] https://live.european-language-grid.eu/catalogue/search/bulgaria?&entity_type__term=Organization
[49] https://www.ontotext.com
[50] This section has been provided by the editors.

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[51] 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
3. **Moderate support**: the language is present in $\geq$10% and $<$30% of the ELG resources of the same type

[51] 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.

4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[52]

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[53] 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.[54]

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

[52] 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.

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

[54] 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.

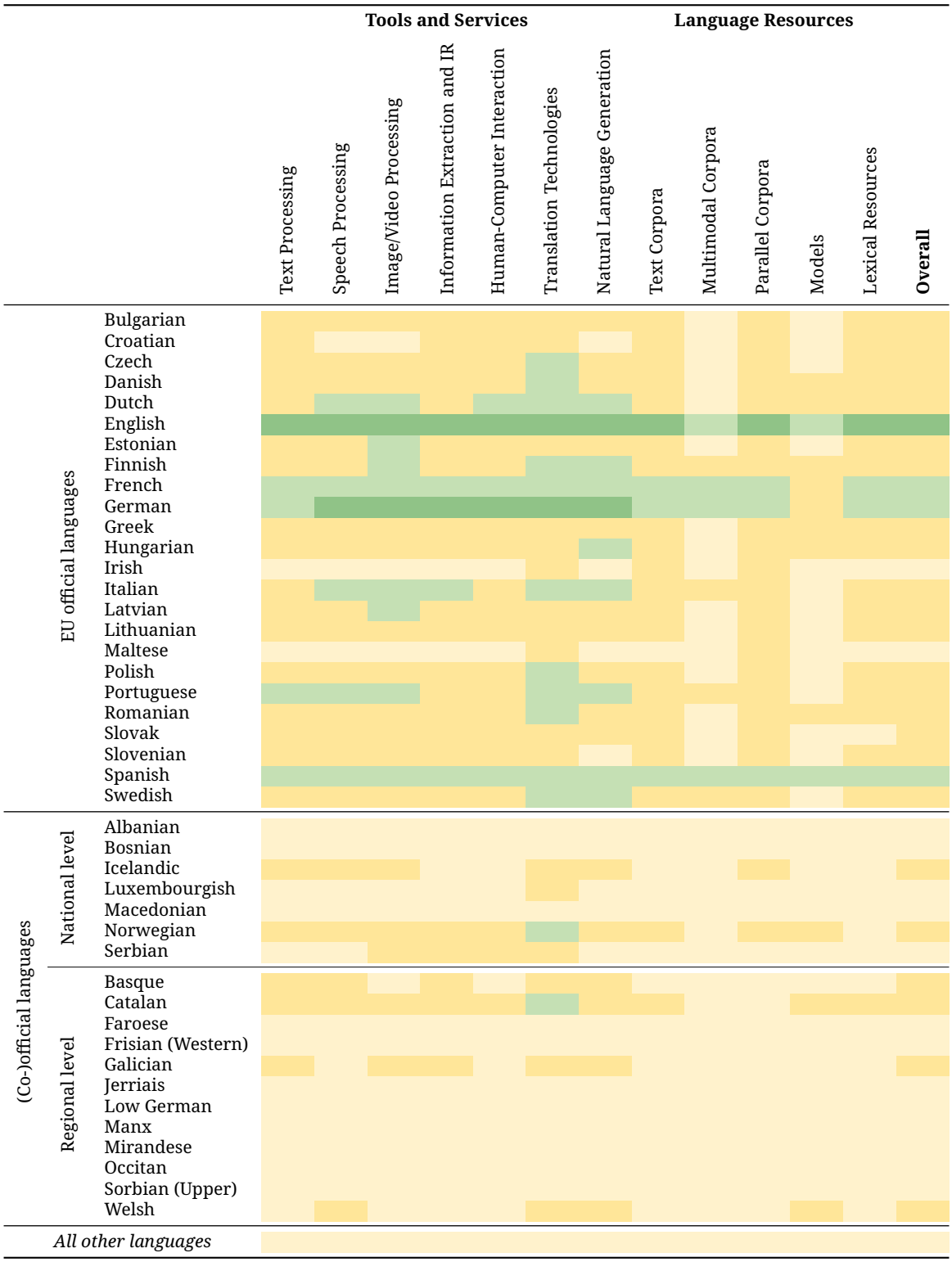

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

national or regional level in at least one European country and other minority and lesser spoken languages,[55] 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.

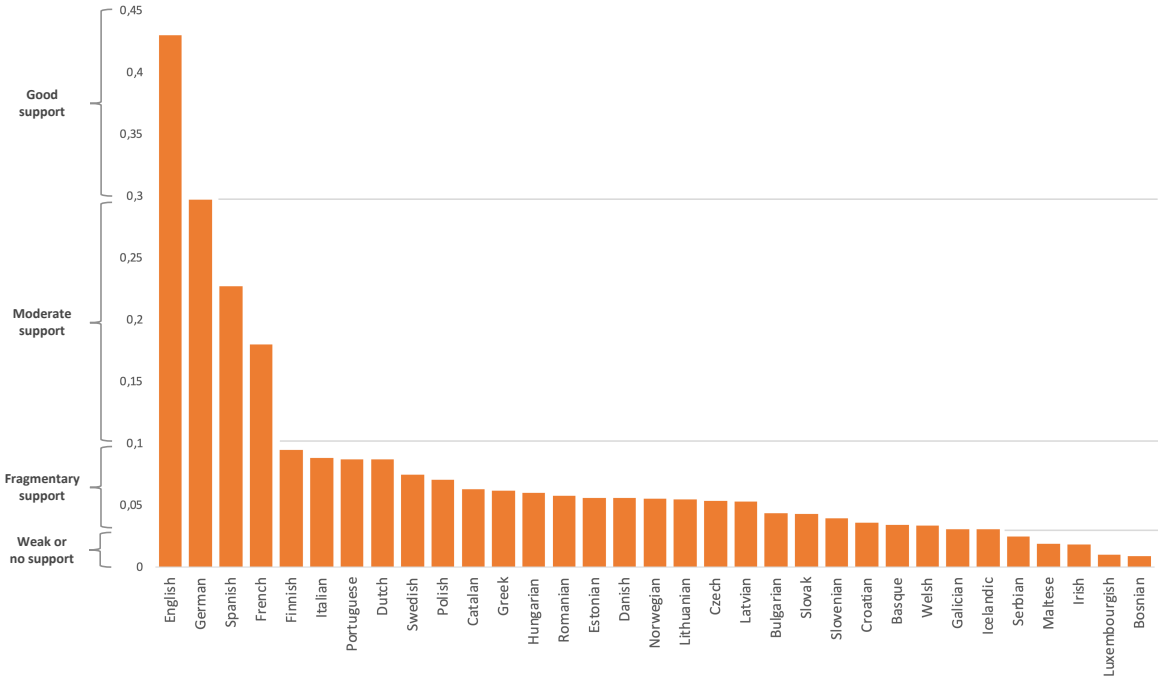

Figure 1: Overall state of technology support for selected European languages (2022)

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 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 direct comparison between the two studies can therefore not be made. However, we can in-

[55] 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.

stead 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

Artificial Intelligence is already a part of everyday life. We use language technologies and Artificial Intelligence when browsing the internet, shopping online, interacting with smart devices and appliances (internet of things), etc. There have been multiple developments in Speech and Language Technology over the past ten years. However, as noted above, many commonly used and necessary technologies are still not available (human-computer interaction, multimodal processing, etc.) and for others, if some advance in technologies is recorded, there are no available applications (summarisation, question answering, etc.). Many technologies are advanced abroad and Bulgarian is part of some multilingual systems for machine translation, speech analysis and recognition.

Change of focus

There has been an increase in focus on the creation of open-source corpora and lexical and conceptual resources. However, not all have been specifically designed for the purpose of developing data-driven LT systems. Furthermore, a shift to intelligent solutions should see a broadening of priorities in terms of funding within the wider lens of Speech and Language Technology (e. g., CALL, speech recognition, human-computer interaction, etc.).

Untapped Potential

There is much untapped, currently inaccessible data that could make a huge impact on the future of Bulgarian LT, if collected and applied appropriately. For example, there are a lot of aligned audio and subtitle text data available in the archives of the national broadcasters (Bulgarian National Television and Bulgarian National Radio) that could be used to build multimodal processing systems.

Need for Dedicated LT Programmes

The focus until now has been on the education in informatics in secondary education and active participation of the IT business in the education in information technologies in schools. There are two master programs in Computational Linguistics at Sofia University. However, there is a need of dedicated education and training programmes in the field of Language Technologies and Artificial Intelligence, as it has proven difficult to source researchers, linguists or engineers with the right combination of skills (e. g. Bulgarian language, computer

science, linguistics). The need is pressing for a new vision and conceptual change of the educational system, lifelong learning and retraining in the field of Language Technologies and Artificial Intelligence.

Long-term strategy

In the absence of a Digital Language Strategy, there are no long-term funding schemes or research centres dedicated to Bulgarian LT. A change in this regard will ensure: support for dedicated LT education and training, investments in data collection and annotation, development of sophisticated LT tools and services that are production ready and easily integrated into smart phone or online applications. A long-term strategy is also related to the existence of solid European support for the development of LT and AI through the provision of strategic and program documents, targeted funding, pan-European cooperation, basic legal framework and ethical norms, and the transfer of good practices.

Open-source culture

As noted above, there are many high quality resources available for Bulgarian that are under strict copyright protection, rendering them unusable for general purposes. It is important therefore, where possible, that all data and tools developed for Bulgarian are made available. This will ensure that their use is not restricted to small institutions that might not have the skills or resources to further develop them for application and general use.

Collaboration

Finally, the benefits of internal and international collaborations with other research centres and industry partners in the advances of LT for Bulgarian are seen. In particular, the EU-funded CEF projects allowed for the sharing of resources in order to reach data collection for Machine Translation development goals.

Vision

Although a number of technologies and resources for Bulgarian exist, there are fewer technologies and resources for the Bulgarian language than for the English language and some other European languages such as German, French, Italian, Spanish. The vision for the future is high-quality Language Technology for all European languages that supports political and economic unity through cultural diversity.

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.

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. Tuning multilingual Transformers for language-specific Named Entity Recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3712. URL https://aclanthology.org/W19-3712.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality*, 11(3):1–27, Jul 2019. ISSN 1936-1963. doi: 10.1145/3297722. URL http://dx.doi.org/10.1145/3297722.

Tiberiu Boroș, Stefan Daniel Dumitrescu, and Ruxandra Burtica. NLP-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K18-2017.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Yoan Dinkov, Ivan Koychev, and Preslav Nakov. Detecting Toxicity in News Articles: Application to Bulgarian. *arXiv preprint arXiv:1908.09785*, 2019.

Gloria Hristova. Text analytics in Bulgarian: An overview and future directions. *Cybernetics and Information Technologies*, 21(3):3–23, 2021. doi: doi:10.2478/cait-2021-0027. URL https://doi.org/10.2478/cait-2021-0027.

Borislav Kapukaranov and Preslav Nakov. Fine-grained sentiment analysis for movie reviews in Bulgarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 266–274, Hissar, Bulgaria, September 2015. INCOMA Ltd. URL https://aclanthology.org/R15-1036.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. We built a fake news & click-bait filter: What happened next will blow your mind! *CoRR*, abs/1803.03786, 2018. URL http://arxiv.org/abs/1803.03786.

Svetla Koeva. Language Situation in Bulgaria. In *National language institutions and national languages. Contributions to the EFNIL Conference 2017 in Mannheim*, pages 183–193, Budapest, 2018. Research Institute for Linguistics, Hungarian Academy of Sciences. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Svetla Koeva. Multilingual image corpus: Annotation protocol. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 701–707, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-main.80.

Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. Natural language processing pipeline to annotate Bulgarian legislative documents. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6988–6994, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.863.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Petar Mitankin, Stoyan Mihov, and Tinko Tinchev. Large vocabulary continuous speech recognition for bulgarian. In *Recent Advances in Natural Language Processing, RANLP 2009, 14-16 September, 2009, Borovets, Bulgaria*, pages 246–250, 2009. URL https://aclanthology.org/R09-1046/.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

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.

Milan Straka and Jana Straková. Tokenizing, POS tagging, lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K/K17/K17-3009.pdf.

Hristo Tanev and Josef Steinberger. Semi-automatic acquisition of lexical resources and grammars for event extraction in Bulgarian and Czech. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 110–118, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2416.

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.

Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadic, Bálint Sass, Bartlomiej Niton, Maciej Ogrodniczuk, Piotr Pezik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Florian Pais, Dan Tufis, Radovan Garabík, Simon Krek, Andraz Repar, Matjaz Rihtar, and Janez Brank. The MARCELL legislative corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3761–3768, 2020. URL https://aclanthology.org/2020.lrec-1.464/.

Boris Velichkov, Ivan Koychev, and Svetla Boytcheva. Deep learning contextual models for prediction of sport event outcome from sportsman's interviews. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Varna, Bulgaria, September 2019. INCOMA Ltd. doi: 10.26615/978-954-452-056-4_142. URL https://aclanthology.org/R19-1142.

Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. Apache Spark: A unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, November 2016. ISSN 0001-0782. doi: 10.1145/2934664.

Ваня Мичева, Ана Кочева, Славия Бърлиева, Васил Николов, and Кирил Топалов. *Българските азбуки*. Издателство на БАН „Проф. Марин Дринов", 2021. ISBN 978-619-245-155-4.

Петя Осенова and Кирил Симов. *Формална граматика на българския език*. Институт за паралелна обработка на информацията, Българска академия на науките, 2007. ISBN 978-954-92148-2-6.