EUROPEAN LANGUAGE EQUALITY

D1.29

Report on the Romanian Language

Authors	Vasile Păiș, Dan Tufiș
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	D1.29
Deliverable title	Report on the Romanian Language
Type	Report
Number of pages	22
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	Vasile Păiș, Dan Tufiș
Reviewers	Maria Giagkou, Guðrún Gísladóttir
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

1 Introduction **3**

2 The Romanian Language in the Digital Age **3**
2.1 General Facts 3
2.2 Romanian in the Digital Sphere 4

3 What is Language Technology? **5**

4 Language Technology for Romanian **7**
4.1 Language Data and Tools 7
4.2 Projects, Initiatives, Stakeholders 9

5 Cross-Language Comparison **10**
5.1 Dimensions and Types of Resources 11
5.2 Levels of Technology Support 11
5.3 European Language Grid as Ground Truth 12
5.4 Results and Findings 12

6 Summary and Conclusions **15**

List of Figures

1 Overall state of technology support for selected European languages (2022) . . 14

List of Tables

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) 13

List of Acronyms

AI	Artificial Intelligence
ASRO	Romanian National Standardisation Body
ccTLD	country code top-level domain
CEF-AT	Connecting Europe Facility Automated Translation
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CoRoLa	Contemporary Romanian Language
DSI	Digital Service Infrastructure
EC	European Commission
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
GDPR	General Data Protection Regulation
GPU	Graphisc Processing Unit
HPC	High-Performance Computing
IIT	Institute for Computer Science
LR	Language Resources/Resources
LT	Language Technology/Technologies
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NLP	Natural Language Processing
NLG	Natural Language Generation
RACAI	Romanian Academy Research Institute for Artificial Intelligence
SR	Speaker Recognition
TEU	Treaty on European Union
UAIC	“Alexandru Ioan Cuza” University of Iași
UPB	Polytechnic University of Bucharest
UTCN	Technical University of Cluj-Napoca
SR	Speaker Recognition

Abstract

Natural language processing aims to provide computers with the ability to understand and produce text and spoken words in the same way that humans do. This is an important step towards developing intelligent systems and human-machine interfaces. In this context, 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.

This study reports on the state-of-play as far as Language-centric AI for Romanian is concerned. From the previous META-NET report (Trandabăț et al., 2012b) there have been significant improvements (e. g. creation of a large Romanian national corpus – CoRoLa, a boost in speech technology, a steady progress in written language technologies including machine translation, construction of a national portal for language resources and tools for the Romanian language – RELATE, etc.), but things are far from what they should be. Support for LT and AI through national programmes is still modest, although there are signs of a more active involvement of the policy makers in the strategic planing and funding programs in this domain.

With the advent of deep learning techniques, we are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to structures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. Research in this field is still required in order to produce complex language models, able to capture the characteristics of the Romanian language. Furthermore, large Language Resources (LR) need to be created so that AI systems are able to learn from them.

The European Language Equality (ELE) project helped to identify existing Romanian LRs and index them in the European Language Grid (ELG). This allowed us to compare the number and type of resources across languages. Results indicate that the number of available Romanian resources is less than 10% of corresponding English resources. And, as detailed in Section 5 – Cross Language Comparison, the Romanian language has fragmentary support, as do many of the European languages.

Language Resources must be created according to national and international regulations, such as copyright laws and privacy-preserving requirements. This usually involves agreements with content producers and making use of anonymisation techniques. In addition, both tools and resources should be available for different use cases (research and commercially). Our investigation revealed that currently only approximately 30% of the known Romanian tools and resources are available free of charge for all uses, which is a very small number compared to English.

Bridging the gap between the available Romanian language resources and those for other languages requires the involvement of policy makers in order to create a research agenda and a national plan for the development of such resources (such plans are currently available in other EU countries). Moreover, funding for the development of state-of-the-art language technologies for the Romanian language is needed.

Rezumat

Procesarea limbajului natural urmărește să ofere calculatoarelor posibilitatea înțelegerii și producerii de text și voce, similar felului în care oamenii folosesc limbajul scris și vorbit. Acesta este un prim pas important spre dezvoltarea de sisteme inteligente și interfețe avansate om-mașină. În acest context, Tehnologia Limbajului (Language Technology – LT) reprezintă domeniul științific multidisciplinar care studiază sisteme capabile să proceseze, să

analizeze, să producă și să înțeleagă limbile umane, indiferent de modul de utilizare: scris sau vorbit.

Acest raport prezintă sumar starea de lucruri în România în domeniul Inteligenței Artificiale centrate pe Tehnologiile Limbajului. Față de raportul precedent METANET (Trandabăț et al., 2012b) s-au înregistrat progrese semnificative (de exemplu a fost creat un Corpus de referința pentru limba română de mari dimensiuni – CoRoLa, tehnologiile limbajului vorbit au progresat simțitor, progresul în domeniul tehnologiilor limbajului, inclusiv în traducerea automată, a fost constant, a fost creat un portal pentru resurse și instrumente de prelucrare a limbii române – RELATE, etc.) dar situația este încă departe de ceea ce ar trebui să fie. Sprijinul autorităților pentru tehnologiile limbajului și inteligența artificială este încă modest, deși există semne (promisiuni) privind o implicare mai activă a guvernului în planificarea strategică și programele de finanțare în acest domeniu.

Ca urmare a apariției arhitecturilor de tip ”învățare profundă” (”deep learning”), a fost înregistrată o tranziție de la sisteme de tip ”flux de prelucrare” (”pipeline”), utilizând multiple componente înlănțuite, spre sisteme monobloc, bazate pe rețele neuronale complexe. Acestea reprezintă o direcție activă de cercetare pentru a produce modele de limbă cu rezultate superioare pentru limba română. De asemenea, antrenarea modelelor utilizând sisteme neuronale complexe necesită cantități uriașe de date (text, audio, multimodale). Acestea sunt cunoscute sub denumirea de Resurse Lingvistice (Language Resources – LR).

Proiectul European Language Equality (ELE) a ajutat la identificarea resurselor și uneltelor existente pentru limba română, precum și la indexarea acestora în platforma European Language Grid (ELG). Acest lucru a permis realizarea unei comparații între numărul și tipul de resurse pentru limba română și cele pentru prelucrarea limbii engleze și a altor limbi europene. Rezultatele arată că resursele românești disponibile reprezintă mai puțin de 10% din cele existente pentru limba engleză. Așa cum se poate observa din Secțiunea 5 – comparație cu celelalte limbi europene, limba română are un suport fragmentar, ca majoritatea limbilor europene.

Resursele lingvistice trebuie create cu respectarea normelor naționale și internaționale, cum ar fi legislația care reglementează copyright-ul și dreptul la viața privată. Sunt astfel necesare înțelegeri scrise cu producătorii de conținut precum și utilizarea unor tehnici, manuale sau automate, pentru anonimizarea conținutului. Aceste măsuri îngreunează procesul de creare a unor resurse mari de limbă română. Un ajutor din partea factorilor de decizie ar fi necesar pentru introducerea în legislație de prevederi speciale pentru utilizarea în scop de cercetare a resurselor disponibile în mediul online. De asemenea, pentru a ușura utilizarea sistemelor automate de prelucrare a limbii române, uneltele și resursele de limbă ar trebui să fie disponibile pentru diferite scopuri, atât de cercetare cât și comerciale. Studiul realizat arată că doar 30% din resursele românești sunt disponibile gratuit pentru orice scop, ceea ce este un număr extrem de mic comparativ cu resursele în limba engleză.

Reducerea diferențelor între capabilitățile de prelucrare a limbii române și alte limbi, raportat la numărul de unelte și resurse disponibile, se poate realiza prin implicarea factorilor de decizie în susținerea unor programe de cercetare axate pe tehnologiile limbajului, precum și crearea unui plan național de cercetare în acest domeniu (similar celor existente în alte țări ale Uniunii Europene). Trebuie avut în vedere că nu este suficientă preluarea unor unelte dezvoltate pentru alte limbi și aplicarea lor pe limba română. Având în vedere specificul limbii române (diacritice, semne de punctuație, expresii, reguli gramaticale) este necesară cercetarea și dezvoltarea unor unelte dedicate limbii române.

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 – identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

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

The report has been developed in the frame of the European Language Equality (ELE) project. With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initiatives, the ELE project develops a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

2 The Romanian Language in the Digital Age

2.1 General Facts

The Romanian language which is an official language of the EU is also the official language of Romania. It is spoken by 19.4 million people[2] in Romania and by approximately 3.5 million people[3] in Moldova, where it is unofficially known as a Moldavian language. Speakers of Romanian in other European countries (Albania, Bulgaria, Croatia, Greece, Hungary, North Macedonia, Serbia, Ukraine and others) and communities of immigrants in Australia, Canada, Israel, Latin America, Turkey, USA and Asian countries totals around 4,000,000 Romanian native speakers.[4]

Romanian is an official language in the European Union and in the Autonomous Province of Vojvodina in Serbia. It is one of the languages spoken in the autonomous Mount Athos in Greece, and it was one of the official languages of the Latin Union. It is also a recognised minority language in Ukraine (Trandabăț et al., 2012a). Furthermore, Romanian has four dialects (Sala, 2006): Daco–Romanian, Aromanian (spoken by approximately 500.000 speakers in Albania, Bulgaria, Greece and North Macedonia), Istro–Romanian (15,000 speakers in 2 small areas in the Istrian Peninsula, Croatia) and Megleno–Romanian (about 5,000 speakers in Greece and North Macedonia). Because of their small number of speakers, these dialects are included in the UNESCO Atlas of the World's Languages in Danger (Moseley, 2010).[5] Both the Istro-Romanian[6] and the Megleno-Romanian[7] are marked as "severely endangered" languages, which means that the language is spoken by grandparents and older generations; while the parent generation may understand it, they are not using it for communicating with their children or between themselves.

[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://ec.europa.eu/eurostat/documents/10186/10994376/RO-EN.pdf
[3] https://ec.europa.eu/eurostat/documents/4031688/9684146/KS-01-19-056-EN-N.pdf/c3f8811c-3793-48aa-befa-b8ad753f1131
[4] https://en.wikipedia.org/wiki/Romanian_diaspora
[5] http://www.unesco.org/languages-atlas/index.php?hl=en&page=atlasmap
[6] http://www.unesco.org/languages-atlas/en/atlasmap/language-id-364.html
[7] http://www.unesco.org/languages-atlas/en/atlasmap/language-id-388.html

The Romanian alphabet is based on the Latin script with five additional letters using diacritics Ă, Â, Î, Ș, Ț (with their corresponding lowercase forms: ă, â, î, ș, ț). For the letters Ș and Ț (lowercase forms: ș, ț), two variants have circulated: one with a comma under the letter, and another one with a cedilla. However, only the former is recommended nowadays by the Romanian National Standardisation Body (ASRO), corresponding to unicode characters U+0x218 (Ș), U+0x21A (Ț), U+0x219 (ș) and U+0x21B (ț). Many electronic texts are not written with diacritics. In order to automatically introduce diacritics, programs have been created to recover them in such texts (Nuţu et al., 2019; Tufiș and Ceaușu, 2008; Tufiș and Chițu, 1999). The quotation marks use double low (left) and right marks („ and ”, respectively). However, especially in electronic texts, the ASCII quotation mark character may be encountered (this is a different character from the right quotation mark, which is the unicode character U+0x201D). Dialogues are introduced using quotation dashes (-). The Oxford comma, used in certain English language documents, is considered incorrect in the Romanian language. In titles, only the first letter of the first word is capitalised, the rest of the title making use of regular sentence capitalisation. Names of months and days, as well as adjectives derived from proper names are not capitalised, e. g. februarie (February), vineri (Friday), italian (Italian).

2.2 Romanian in the Digital Sphere

In 2019, 84% of Romanian households had Internet access.[8] This represents an increase of 23% compared to 2014. The proportion of households in rural areas with Internet access is lower than the equivalent proportions of households in cities or in towns and suburbs. According to a EuroStat report,[9] the divide between rural areas and the two other types of areas was particularly strong in Greece, Bulgaria, Portugal, Slovenia and Romania, each of which had a lower overall level of Internet access than the EU-27 average. Furthermore, people who used the Internet in a three-month period accounts for 77%, according to the same report, and out of these users around 60% are daily Internet users. Moreover, around 70% of Internet users are using the Internet on a portable computer or handheld device via a mobile or wireless connection.

With respect to the digital readiness of Romania and the presence of the Romanian language in the digital sphere, around 60% of Romanian Internet users also participated in social media interactions in 2019. The most widely used platform is Facebook.[10] This is followed by Instagram, Twitter and LinkedIn. Other social media platforms are used as well, but in lower numbers. Around 32% of Romanian businesses also make use of social networks.[11] Most of the enterprises (25%) make use of a single social media type. Besides social networks, blogs or micro-blogs are used by 4% of Romanian enterprises. Artists are also using social media platforms to engage with their fans.[12] In this case, the most widely used platform is still Facebook, followed by Instagram and YouTube. Communication in social media is usually done in Romanian. However, code mixing (the interleaving of two or more languages within a sentence or discourse (Belazi et al., 1994)) is a phenomenon encountered in social media posts. In this case, we usually see a mix of Romanian and English words or expressions within the same post.

The Internet country code top-level domain (ccTLD) for Romania is “.ro”. It is administered

[8] https://www.statista.com/statistics/377760/household-internet-access-in-romania/

[9] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Digital_economy_and_society_statistics_-_households_and_individuals

[10] https://www.statista.com/topics/7134/social-media-usage-in-romania/

[11] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Social_media_-_statistics_on_the_use_by_enterprises

[12] https://www.iqads.ro/articol/55181/social-media-stars-index-realizat-de-starcom-romania-pentru-luna-mai-inna-pe

by the National Institute for R&D in Informatics. In 2018, there were 789,833 Romanian top-level domains registered.[13] Throughout the years, there has been a continuous increase in domains registered under the ".ro" top-level domain.[14] Following the European digital strategy[15] and believing in a pan-European Internet identity, a number of companies prefer to register their domains under the ".eu" top-level domain. The language being used on ".ro" top-level domains is usually Romanian, but certain companies prefer to present their message also in English or other languages, depending on the target audience.

3 What is Language Technology?

Natural language[16] 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 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 replaced by data-based ones, i. e. systems that learn implicitly from examples. In the recent decade of 2010s we observed a radical technological change in NLP: the use of multilayer neural networks able to solve various sequential labelling problems. The success of this approach lies in the ability of neural networks to learn continuous vector representations of the words (or word embeddings) using vast amounts of unlabelled data and using only some labelled data for fine-tuning.

In recent years, the LT community has been witnessing the emergence of powerful new deep learning techniques and tools that are revolutionising the way in which LT tasks are approached. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. The

[13] https://mxhost.ro/rotld-domenii.pdf
[14] https://en.wikipedia.org/wiki/.ro
[15] https://digital-strategy.ec.europa.eu/en/policies/eu-top-level-domain
[16] 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).

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 users to converse with computers using natural language (text, speech and non-verbal communication signals, such as gestures and facial expressions). Popular applications within this area are conversational agents (better known as chatbots).

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

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[17]

[17] 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-

4 Language Technology for Romanian

Human language comes in spoken and written forms. While speech is the oldest and in terms of human evolution the most natural form of language communication, complex information and most of human knowledge is stored and transmitted through texts. Speech and text technologies process or produce these different forms of language, using dictionaries, grammar rules, and semantics. This means that Language Technology (LT) links language to various forms of knowledge, independently of the medium (speech or text) in which it is expressed.

When we communicate, we combine language with other modes of communication and information media. For example, speaking can involve gestures and facial expressions. Digital texts link to pictures and sounds. Movies may contain language in spoken and written form. In other words, speech and text technologies overlap and interact with other multimodal communication and multimedia technologies. Even more, with the increased usage of social media, we see the large-scale usage of memes.[18] These usually represent amusing or interesting items (such as captioned pictures or videos) that are spread widely online especially through social media.

4.1 Language Data and Tools

The availability of language-specific data has a direct impact on the quality of language-specific or cross-language tools. This is particularly important for modern tools developed using deep neural networks, which require large amounts of data for training. Availability of large pre-trained multilingual language models that include representations for Romanian language, such as XLM-RoBERTa (Conneau et al., 2020) or mBERT (Devlin et al., 2019), somewhat alleviate the problem of constructing computationally intensive contextual word representations. Nevertheless, monolingual representations usually lead to increased performance of monolingual tools. In this context, different initiatives lead to the construction of large Romanian contextual models, like RoBERT (Masala et al., 2020), Romanian BERT (Dumitrescu et al., 2020) and even domain-specific models, like jurBERT (Masala et al., 2021). Nevertheless, static word representations, such as CoRoLa-based word embeddings (Păiș and Tufiș, 2018) and others, are largely used for training different tools due to lower computing requirements.

Word representations, either contextual or static, form only the basis of advanced language tools. In addition to these language models, additional task-specific corpora is required to train and evaluate the tools. As a direct result of the Language Equality Project (ELE),[19] available language data was investigated and indexed by the European Language Grid (ELG).[20] Out of the identified Romanian Language Resources (LRs), the vast majority are multilingual, some are bilingual and only a few are monolingual corpora. Compared to the English language, the available Romanian corpora represents 9.11%. When it comes to monolingual corpora, the difference is even greater, the available Romanian monolingual corpora representing only 3.8% of the available English monolingual corpora. Considering the neighboring EU member states in the region, the number of Romanian corpora is the lowest (when comparing with Hungarian and Bulgarian). The situation remains unchanged

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).

[18] https://www.merriam-webster.com/dictionary/meme

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

[20] https://www.european-language-grid.eu

when considering monolingual corpora. Romanian lexical and conceptual resources represent approximately 10% of the available English resources, while grammar and language models represent 20%. Considering neighboring EU member states, Romanian resources of this kind are near the region's average. The available speech corpora containing Romanian audio represent 5% of the available English resources and approximately 50% when compared to Hungarian and Bulgarian resources. Overall, when compared to English, available Romanian resources seem insufficient.

In spite of the reduced number of available language resources, applications for different natural language processing tasks were constructed for Romanian language. This includes lemmatisation, part-of-speech tagging, dependency parsing, named entity recognition, syllabification, automatic speech recognition, text-to-speech, automatic translation, punctuation restoration, terminology annotation, text classification. Following the ELE project, over 100 tools and services for Romanian language were identified. This number represents 15% of the available English tools. Looking at language tools from neighboring EU member states, we find a similar number for Bulgarian and slightly higher for Hungarian tools. Out of the available Romanian tools, approximately 30% allow processing of audio input data and 10% of the tools allow for audio output.

Considering the availability of the tools and resources from a licensing perspective, we were able to find that aproximately 30% of the tools are available without a fee for all uses with an additional 10% available without a fee for specific uses. Similarly, we were able to identify approximately 30% of the language resources as being available without a fee for all uses with an additional 10% available without fee only for specific uses. Thus, we conclude that most resources are available for specific uses only, usually for research purposes, or for a fee, thus limiting their value for innovation.

Even if, in general, all language technology fields are covered, there are certain fields that are less developed or not yet considered for the Romanian language by researchers and developers: language generation, dialogue management systems, multimodal corpora building, social media aspects (including messages, micro-blogging, social networks, memes interpretation). Speech processing is currently much less mature than language technology for written text, both in terms of corpora and instruments. Even though there has been much work on processing general Romanian language, more focus is needed for creating domain-specific language resources and tools (especially for the biomedical, legal, economy and social media domains).

A legally unclear situation restricts the usage of digital texts, such as those published online by newspapers, for empirical linguistics and language technology research, for example, to train statistical language models. Together with politicians and policy makers, researchers should try to establish laws or regulations that enable researchers to use publicly available texts for language-related R&D activities. This situation is made more complicated by privacy-preserving requirements, deriving from the GDPR and other regulations. In this context, mature and robust anonymisation technologies are required, considering domain-specific needs (for example anonymisation requirements related to medical or legal documents). In this context, the Representative Corpus of Contemporary Romanian Language (CoRoLa)[21] (Barbu Mititelu et al., 2019) was created in a priority project of the Romanian Academy as the largest IPR-cleared reference corpus of written and spoken Romanian. Texts cover 4 domains (arts and culture, science, society, nature) organised in 70 subdomains, reflecting 6 styles (imaginative, journalistic, scientific, legal, administrative, memoirs) and different document types (entire books, book chapters, newspaper/magazine articles, scientific articles, Wikipedia articles, news, interviews, blog posts, letters, reports, etc.).

One of the largest Romanian read speech corpus is RSC (Georgescu et al., 2020), containing 100 hours of recorded audio files. The multilingual speech corpus VoxPopuli (Wang et al.,

[21] https://corola.racai.ro

2021) contains 83 hours of Romanian language speech. The speech component of the CoRoLa corpus (comprised of multiple smaller corpora together with additional audio files specifically obtained for inclusion in CoRoLa) a total number of 103 hours of sound aligned with the corresponding text.

4.2 Projects, Initiatives, Stakeholders

A number of Romanian language technologies, covering different fields of research, are available within the RELATE[22] (Păiș et al., 2020) portal. This integration allows for direct usage of tools developed at the Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy and by partners in different research projects. The platform covers results derived from more than 6 national and international research projects. As part of the integration effort within the platform, we were faced with the different formats for both data and APIs. Since there is no standardisation with regard to data format, different institutions employed specific formats, considered at the time to be more easy to use for specific needs. However, this approach makes it more difficult to integrate resulting resources and tools into a unifying platform. Therefore, we consider that a concerted programme is required to standardise data formats and APIs, in order to allow their re-use and integration into complex applications.

As already mentioned in the D3.1 report of the ELE project (Aldabe et al., 2021), the last roadmap from the European Strategy Forum on Research Infrastructures (ESFRI) includes Big Data technology as one of the emerging drivers of the landscape analysis. Regarding LT, the ESFRI Landmark CLARIN ERIC (Common Language Resources and Technology Infrastructure) offers interoperable access to language resources and technologies for researchers in the humanities and social sciences.[23] Unfortunately, not all EU Member States are official members of CLARIN. This includes Romania, who is not a member of CLARIN. CLARIN offers access to language data, tools to work with the data, and expertise about such resources. Several services and tools can be used online even without downloading the data. This is achieved by several components or services, such as the CLARIN portal, discovery tools, federated identity, virtual collections, persistent identifiers, workspaces, online tool chains and many more individual services at the many CLARIN centres. According to the federated identity concept, it is possible to access the CLARIN centres with one's own (academic) credentials, based on a trust network of academic organisations. Thus, CLARIN is working at crossing the country borders when accessing resources, so that e. g. a Romanian researcher (if Romania were a member) could access language resources hosted in Austria or Italy, relevant to a specific multilingual research project.

Regarding AI, various documents have been published recently by the European institutions: *European AI leadership, the path for an integrated vision*,[24] the Strategy on AI,[25] the *Ethics Guidelines for Trustworthy AI*,[26] *Liability for AI and other emerging technologies*,[27] the *White Paper on AI*,[28] and the *Coordinated Plan on AI*.[29] They all agree that AI is an area of strategic importance and key driver of economic development and that it can provide solutions to many societal challenges. In this context, many EU countries also have national plans

[22] https://relate.racai.ro
[23] http://www.clarin.eu
[24] https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2018)626074
[25] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN
[26] https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf
[27] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=63199
[28] https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
[29] https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review

for AI (for example the *Spanish National AI Strategy*[30] or the *French AI for Humanity*[31]). In Romania however there is currently no such national plan for AI or for language technologies. A strategy for AI[32] has been proposed recently within the RePatriot project,[33] but this was not adopted at national level. Furthermore, the AI Strategy is not very concrete. It centers mostly on which Romanian sectors would benefit from AI, and which steps are important in the process of developing and implementing Romanian AI initiatives, but it does not include any plans about how to accomplish these actions.

Several Romanian universities include AI related classes in their curricula and have master and doctoral programs in the field of AI and language technologies. In both research institutes and universities, research is being conducted for developing and applying AI techniques to processing Romanian language and creating new language resources. Important research centers focusing on LT and LR can be found in the Romanian Academy Research Institute for Artificial Intelligence (RACAI), Institute for Computer Science (IIT)), the Polytechnic University of Bucharest (UPB), the "Alexandru Ioan Cuza" University of Iași (UAIC), the Technical University of Cluj-Napoca (UTCN). This is not an exhaustive list, since language research, at different levels, is conducted in other institutes and universities as well. Furthermore, industry awareness of AI in general and language technologies in particular has increased. This led to the emergence of Romanian companies offering AI and Romanian language processing solutions.

In the recent ReTeRom project,[34] which involved a collaboration between RACAI, UAIC, UPB and UTCN, multiple Romanian resources and tools were developed. The RoLEX lexicon[35] is a resource with 330,866 entries for syllabification and phonetic transcription of words. It is the largest of its kind for Romanian language. TEPROLIN (Ion, 2018) is a web service able to perform 15 text processing operations for Romanian. It was further integrated in the RELATE platform. The Connecting Europe Facility – Automated Translation (CEF-AT) projects MARCELL[36] and CURLICAT[37] provide large corpora in 7 EU languages, including Romanian, aiming to enhance Machine Translation (MT) capabilities. MARCELL focused on national legislation (Váradi et al., 2020; Tufiș et al., 2020), while CURLICAT focuses on domains of relevance to all the European Digital Service Infrastructures (DSIs).

5 Cross-Language Comparison

The LT field[38] 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).

[30] https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/021220-ENIA.pdf
[31] https://www.aiforhumanity.fr/en/
[32] https://www.slideshare.net/MonicaIon1/strategy-romania-in-the-era-of-artificial-intelligence-rblrepatriot
[33] https://repatriot.ro
[34] https://www.racai.ro/p/reterom/index_en.html
[35] https://www.racai.ro/p/reterom/results.html
[36] https://marcell-project.eu
[37] https://curlicat-project.eu
[38] 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[39] 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

[39] 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[40]

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[41] 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.[42]

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

[40] 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.

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

[42] 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.

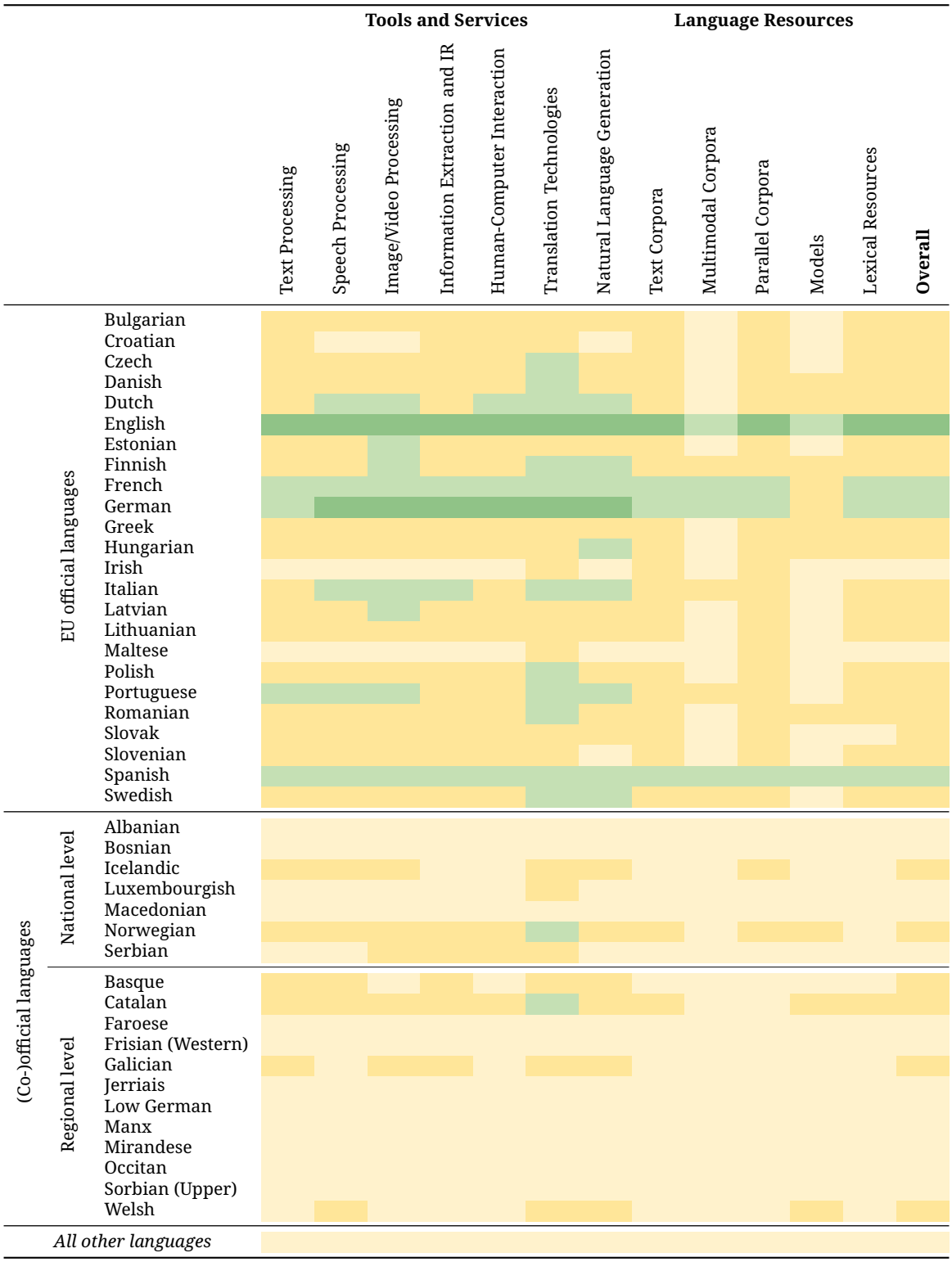

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

national or regional level in at least one European country and other minority and lesser spoken languages,[43] 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.

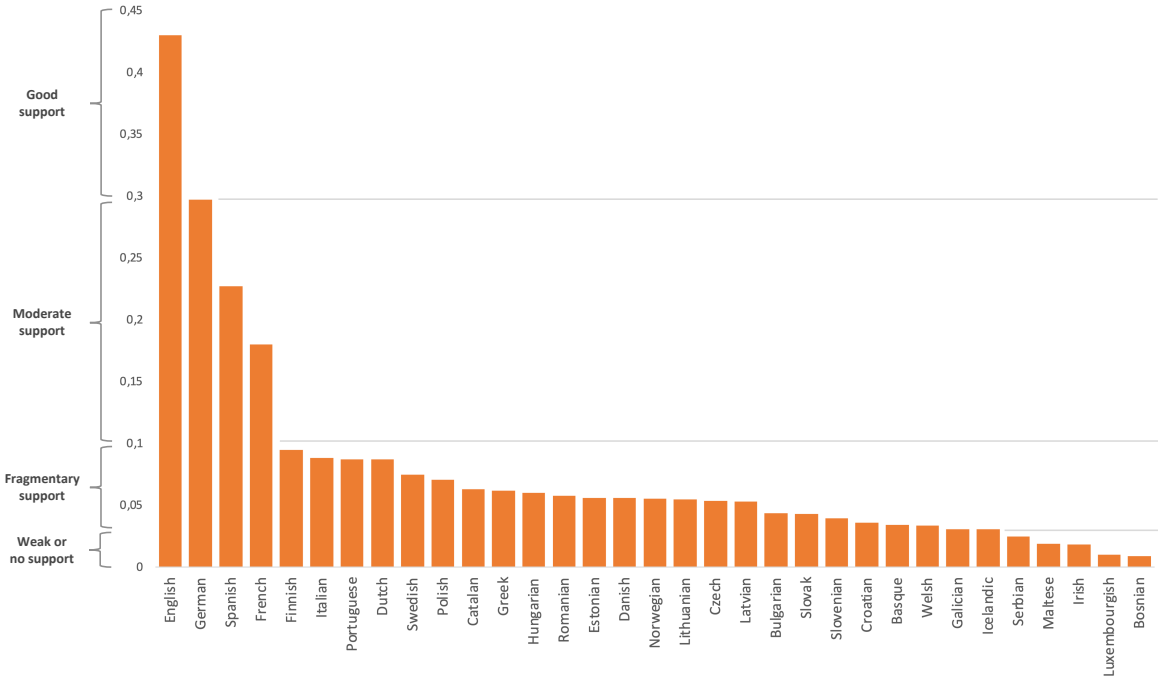

Figure 1: Overall state of technology support for selected European languages (2022)

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-

[43] 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

This report continues the efforts started in 2012 with the META-NET White Paper series to assess the Language Technology support for European languages and to provide a high-level comparison across these languages. Much has happened in the past 20 years: the number of Romanian households with Internet access increased by over 20% (Section 2), language technologies have become fused in our everyday lives (Section 3), large-scale language models with support for Romanian language have become available (Section 4.1), and many research institutes and universities are participating in the development of Romanian language resources (Section 4.2).

Even though progress has been made, there is still a huge discrepancy between the number of available resources for the Romanian language and those available for English or other European languages. This has a direct influence on the quality of tools for analyzing Romanian language. Current state-of-the-art natural language processing systems, based on deep neural network architectures, employ large collections of texts, speech or multimodal data in order to learn how to perform different tasks. Therefore, a reduced number of resources ultimately impact the quality of Romanian language processing systems.

In order to make Romanian language technology competitive with those of other European languages, there is a tremendous need for large-scale linguistic resources, from raw Romanian texts and speech recordings to heavily annotated data (highlighting particular linguistic phenomena). Furthermore, state-of-the-art tools adapted to processing Romanian language must be developed using the newly created resources. One way to achieve these goals is through a dedicated long-term research and development funding programme (similar to national language technology research programmes available in other European countries). Furthermore, in order to facilitate the creation of language resources, content producers (newspapers, publishing houses) should be made aware of the importance of sharing parts of their content for research purposes.

By increasing the quality of tools available for Romanian language (including text translation, speech translation, information extraction) it will help tear down existing language barriers and build bridges between Europe's languages, thus paving the way for political and economic unity through cultural diversity. This is in agreement with the Treaty on European Union (TEU) which explicitly aims to "respect its rich cultural and linguistic diversity and to ensure that Europe's cultural heritage is safeguarded and enhanced" (Article 3 of the TEU). In the present digital age, one way to safeguarding a nation's cultural heritage is through digitisation and by employing AI mechanisms to allow easy indexing and retrieval of information regardless of the language employed.

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.

Verginica Barbu Mititelu, Dan Tufiș, Elena Irimia, Vasile Paiș, Radu Ion, Nils Diewald, Maria Mitrofan, and Onofrei Mihaela. Little strokes fell great oaks. creating CoRoLa, the reference corpus of contemporary Romanian. In *Revue roumaine de linguistique, No./Issue 3*, 2019.

Hedi M. Belazi, Edward J. Rubin, and Almeida Jacqueline Toribio. Code switching and x-bar theory: The functional head constraint. *Linguistic Inquiry*, 25(2):221–237, 2021/11/17/ 1994. URL http://www.jstor.org/stable/4178859. Full publication date: Spring, 1994.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.387. URL https://aclanthology.org/2020.findings-emnlp.387.

Alexandru-Lucian Georgescu, Horia Cucu, Andi Buzo, and Corneliu Burileanu. RSC: A Romanian read speech corpus for automatic speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6606–6612, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.814.

Radu Ion. *TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian*, pages 69–76. Editura Universității "Alexandru Ioan Cuza", Iași, 2018.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. RoBERT – a Romanian BERT model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.581. URL https://aclanthology.org/2020.coling-main.581.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. jurBERT: A Romanian BERT model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.nllp-1.8.

Christopher Moseley, editor. *Atlas of the World's Languages in Danger, 3rd edn.* UNESCO Publishing, 2010. URL http://www.unesco.org/culture/en/endangeredlanguages/atlas.

Maria Nuţu, Beáta Lőrincz, and Adriana Stan. Deep learning for automatic diacritics restoration in Romanian. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 235–240, 2019. doi: 10.1109/ICCP48234.2019.8959557.

Vasile Păiș and Dan Tufiș. Computing distributed representations of words using the CoRoLa corpus. *Proceedings of the Romanian Academy Series A - Mathematics Physics Technical Sciences Information Science*, 19(2):185–191, 2018. URL https://academiaromana.ro/sectii2002/proceedings/doc2018-2/Art12Pais.pdf.

Vasile Păiș, Dan Tufiș, and Radu Ion. A processing platform relating data and tools for Romanian language. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 81–88, Marseille, France, May 2020. European Language Resources Association. URL https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/IWLTP2020book.pdf#page=87.

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.

Marius Sala, editor. *Enciclopedia limbii române (Encyclopaedia of the Romanian Language), ediția a 2-a (2nd Edition).* Univers Enciclopedic, București, 2006.

Diana Trandabăț, Elena Irimia, Verginica Barbu Mititelu, Dan Cristea, and Dan Tufiș. *The Romanian Language in the Digital Age*. Springer Publishing Company, Incorporated, 2012a. ISBN 3642307027.

Diana Trandabăț, Elena Irimia, Verginica Barbu Mititelu, Dan Cristea, and Dan Tufiș. *Limba română în era digitală – The Romanian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012b. ISBN 978-3-642-30702-7. Available online at http://www.meta-net.eu/whitepapers.

Dan Tufiș and Alexandru Ceaușu. *DIAC+: Un sistem profesional de recuperare a diacriticelor*, pages 151–160. Editura Universității "A.I. Cuza", Iași, 2008.

Dan Tufiș and Adrian Chițu. Automatic insertion of diacritics in Romanian texts. In Ferenc Kiefer, Gábor Kiss, and Júlia Pajzs, editors, *Proceedings of the 5th International Workshop on Computational Lexicography (COMPLEX 1999)*, pages 185–194, Pecs, Hungary, may 1999. Linguistics Institute, Hungarian Academy of Sciences. URL http://www.racai.ro/media/Tufis-Chitu-COMPLEX1999.pdf.

Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. Collection and annotation of the Romanian legal corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2766–2770, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.337/.

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 Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Barbu Mititelu, Verginica, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar, and Janez Brank. The marcell legislative corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3754–3761, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.464/.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. URL https://aclanthology.org/2021.acl-long.80.