EUROPEAN LANGUAGE EQUALITY

D1.31

Report on the Slovenian Language

Author	Simon Krek
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	D1.31
Deliverable title	Report on the Slovenian Language
Type	Report
Number of pages	24
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
Author	Simon Krek
Reviewers	Tea Vojtěchová, Maria Eskevich
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

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

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CLARIN.SI	CLARIN Slovenia
CTS	Core Trust Seal
DLE	Digital Language Equality
ELAN	European Language Activity Network
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)
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
EU	European Union
GOS	Govorjena Slovenščina
GPU	Graphics Processing Unit
HPC	High-Performance Computing
IPA	International Phonetic Alphabet
IT	Information Technology
LR	Language Resources/Resources
LT	Language Technology/Technologies
MA	Master of Arts
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
NLG	Natural Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
POS	Part-of-Speech
SAMPA	Speech Assessment Methods Phonetic Alphabet
SOFES	Speech Database of Spoken Flight Information Enquiries
TM	Translation Memory

Abstract

Around 2.5 million people around the world speak or understand Slovene, with a vast majority of them living in the Republic of Slovenia where Slovene is the official language. In certain municipalities, the constitution grants the right to use their mother tongue to Italian and Hungarian minorities. According to legislation in Slovenia, all education and teaching provided as part of the current state curriculum, from pre-school through to university level, must be in Slovene.

While the first written resources identified as Slovene date from the late 10th century, the language was standardised and described for the first time during the Protestant Reformation in the 16th century. In the second half of the 19th century standardisation process was largely concluded, when the "gajica" script was generally accepted. The development of standard Slovene was complicated by the large number of dialects – more than 40 dialects in seven larger dialect groups. Modern standard Slovene is to a large extent still considered as a written standard while spoken Slovene consists of a large variety of spoken idioms determined by region, local dialect, age group, education and other demographic factors. A distinctive feature of Slovene is the existence of dual grammatical number in the declension of nouns, adjectives, pronouns and numerals, as well as in verb conjugation. Slovene is one of the rare Indo-European languages where this feature has survived from the hypothetical Proto-Indo-European language.

In 2021, 93% of households in Slovenia had access to the Internet. Statistics also shows that 85% of people aged between 16 and 74 use the Internet every day or almost every day, and 74% of them ordered or bought goods or services online in the last 12 months. Slovene is used extensively on the web and in social media. Computer-mediated communication differs from standard language variant to a significant degree, which represents an additional challenge to the language technologies.

Historically, widespread language technology applications began with the first Slovene spelling checkers at the beginning of 1990s. The first international and national funding in mid-1990s provided the first Slovene language resources with standard markup and annotation. At the same time, speech technologies began to be funded in international and national projects. Between 2008-2013 Slovene government funded an extensive LT-oriented project (Communication in Slovene) and a major breakthrough was achieved when Slovene CLARIN consortium received stable funding in 2015, and entered into the European CLARIN ERIC infrastructure. All major Slovene institutions involved in the development of LT resources, tools and services are members of the CLARIN.SI consortium.

From 2018, Slovene Ministry of Culture started financing several projects which produced different language resources. However, the biggest investment in LT for Slovene is the "Development of Slovene in Digital Environment" project (2020-2023) which will significantly upgrade existing LT resources, tools and services, or produce those that do not exist yet. In general, in the last ten years the government in Slovenia recognised the need to support LT for Slovene, and language technology entered into national planning documents and funding schemes.

On the other hand, the number of private companies in Slovenia specialising in LT for Slovene remains extremely low, and most of the LT products come either from the (Slovene) academic sphere with national or EU funding, or from the big international IT companies that cover a large number of languages.

Povzetek

Slovenščino govori približno 2,5 milijona ljudi po vsem svetu, od tega jih večina živi v Republiki Sloveniji, kjer je slovenščina ustavno opredeljena kot uradni jezik. V nekaterih delih države je pravica do uporabe maternega jezika priznana tudi italijanski in madžarski manjšini. V skladu z zakonodajo mora biti pouk, ki se izvaja v okviru veljavnega državnega kurikula, od predšolske do univerzitetne ravni, v slovenskem jeziku.

Prvi pisni viri, opredeljeni kot slovenski, izvirajo že s konca 10. stoletja, vendar je bila slovenščina prvič standardizirana in opisana šele v času protestantske reformacije v 16. stoletju. V drugi polovici 19. stoletja se je proces standardizacije večinoma zaključil, ko je bila splošno sprejeta nova pisava "gajica". Razvoj standardne slovenščine je bil zapleten tudi zaradi velikega števila narečij – več kot 40 v sedmih večjih narečnih skupinah. Sodobna standardna slovenščina je zato v precejšnji meri še vedno opredeljena kot pisni standard, medtem ko govorjeno slovenščino sestavlja veliko število individualnih govorov, ki jih določa regija, lokalno narečje, starostna skupina, izobrazba in drugi demografski dejavniki. Posebna značilnost slovenščine je obstoj dvojine kot slovničnega števila pri pregibanju samostalnikov, pridevnikov, zaimkov, številkov in glagolov. Slovenščina je eden redkih indoevropskih jezikov, kjer se je ta značilnost ohranila iz indoevropskega prajezika.

Zgodovinsko gledano se je raba jezikovnih tehnologij za slovenščino začela s prvimi črkovalniki v začetku devetdesetih let prejšnjega stoletja. V okviru mednarodnih in nacionalnih projektov so bili sredi devetdesetih let prejšnjega stoletja izdelani prvi slovenski jezikovnotehnološki pisni viri, hkrati je se začelo tudi financiranje govornih tehnologij. Konec prejšnega stoletja (1998) se je jezikovnotehnološka skupnost organizirala v okviru *Slovenskega društva za jezikovne tehnologije* (1998), ki vsaki dve leti organizira znanstveno konferenco. V letih 2008-2013 je bil financiran obsežen jezikovnotehnološki projekt *Sporazumevanje v slovenskem jeziku*, velik preboj pa je bil dosežen, ko je slovenski konzorcij CLARIN leta 2015 prejel stabilno financiranje in vstopil v evropsko infrastrukturo CLARIN ERIC. V konzorciju CLARIN.SI so zbrane vse pomembnejše slovenske institucije, ki se ukvarjajo z razvojem jezikovnotehnoloških virov, orodij in storitev.

Z letom 2018 je slovensko Ministrstvo za kulturo začelo financirati več projektov, v okviru katerih so nastali različni jezikovni viri, največjo naložbo v jezikovne tehnologije za slovenščino do sedaj pa predstavlja projekt *Razvoj slovenščine v digitalnem okolju* (2020-2023), v okviru katerega bodo bistveno nadgrajeni obstoječi viri, orodja in storitve oziroma izdelani tisti, ki še ne obstajajo. Na splošno velja, da je vlada Republike Slovenije v zadnjih desetih letih prepoznala potrebo po podpori jezikovnim tehnologijam za slovenščino, ta tema je bila upoštevana tudi v nacionalnih načrtovalnih dokumentih in shemah financiranja. Hkrati je treba poudariti, da je število jezikovnotehnoloških podjetij, ki zagotavljajo storitve specifično za slovenščino, še vedno izredno majhno, večina jezikovnotehnoloških izdelkov pa prihaja bodisi iz (slovenske) akademske sfere z nacionalnim ali evropskim financiranjem ali iz velikih mednarodnih IKT podjetij, ki pokrivajo veliko število jezikov.

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.

2 The Slovene Language in the Digital Age

2.1 General Facts

Official status and education

It has been estimated that around 2.5 million people around the world speak or understand Slovene, with a vast majority of them living either in the Republic of Slovenia or in the neighbouring areas in Italy, Austria, Hungary and Croatia. In the national census of 2002 (Statistical Office of the Republic of Slovenia, 2003), the last one that recorded the number of native speakers of different languages in Slovenia, 87.8% of the population – of a total of just under 2 million at the time – declared Slovene to be their mother tongue, with another 3.3% claiming that they use Slovene as the language of their everyday communication at home. This amounts to 91.1% of the population using Slovene as their first language. This number puts Slovenia in the group of EU states with the most homogeneous linguistic situation. Among other linguistic groups, native speakers of languages of former Yugoslavia were the largest in 2002. 3.3% of them used a combination of Slovene and their mother tongue for everyday communication, and another 1% used only their mother tongue – Bosnian, Croatian, Serbian or Montenegrin. Other smaller communities included speakers of Albanian, Macedonian and Romani . Similar to many cases in European history, rather complex developments in the past led to the situation where relatively large Slovene minorities now live in the region Friuli-Venezia Giulia in Italy, in Austrian federal states Kärnten and Steiermark, as well as in the bordering area with Hungary and in Croatian Istria. On the other hand, Italian and Hungarian minorities live in the bordering regions in Slovenia.

Slovene is the official language in the Republic of Slovenia. The constitution grants the right to use their mother tongue to the two minorities declaring that “in those municipalities where Italian or Hungarian national communities reside,” Italian or Hungarian are also official languages. In 2002, it was recorded that Hungarian is the mother tongue of 0.4% of the population, and Italian of 0.2%.

According to legislation in Slovenia, all education and teaching provided as part of the current state curriculum, from pre-school through to university level, must be in Slovene. In pre-school, primary and secondary education, Italian is used in the schools of the Italian minority community, while Hungarian and Slovene are used in bilingual schools where the Hungarian minority is found. Special arrangements exist for children whose mother tongue is not Slovene, for the education of Roma children, children of foreign citizens and children of people without citizenship.

[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

History and grammatical characteristics

While the first written resources identified as Slovene date from the late 10th century, the language was standardised and described for the first time during the Protestant Reformation in the 16th century. In 1550, Protestant reformer Primož Trubar published first two Slovene books "Catechismus" and "Abecedarium". The other two most important Protestant works were the Bible translated into Slovene by Jurij Dalmatin and the Slovene Grammar by Adam Bohorič, both published in 1584. In the second half of the 19th century standardisation process was largely concluded, when the new "gajica" script was generally accepted. The most obvious difference between the previously used "bohoričica" script (named after the first grammar writer Adam Bohorič) and the new one was the replacement of the letters ſ and s by s and z, and letter pairs zh, sh, ſh by the accented letters č, ž, š, used also today in the standard 25-letter Slovene alphabet using Latin script.

In addition to the often precarious political circumstances hindering the use of Slovene in all spheres of life – throughout history the region had been part of larger political entities, usually with centralisation and unilingual tendencies – the development of standard Slovene was further complicated by the large number of dialects. There are now more than 40 dialects recognised in seven larger dialect groups. Modern standard Slovene is therefore, to a large extent, still considered as a written standard while spoken Slovene consists of a large variety of spoken idioms determined by region, local dialect, age group, education and other demographic factors. Regional standards do exist and are used in general public speech; however, the highest form of Slovene pronunciation – the equivalent of Received Pronunciation in English – is predominantly spoken by professionals at the National Radio and Television or on formal occasions.

A distinctive feature of Slovene that has important consequences also for computational processing of natural language is the existence of dual grammatical number in the declension of nouns, adjectives, pronouns and numerals, as well as in verb conjugation. Slovene is one of the rare Indo-European languages where this feature has survived from the hypothetical Proto-Indo-European language. Therefore, in almost all nouns, the dual grammatical number is expressed with different inflections as shown in Table 1.

	singular	**dual**	**plural**
chair (masc.)	stol	stol**a**	stol**i**
table (fem.)	miza	miz**i**	miz**e**
window (neut.)	okno	okn**i**	okn**a**

Figure 1: Dual grammatical number in the declension of nouns

Slovene nouns also show six grammatical cases and three genders with several inflectional paradigms which leads to an explosion of different inflectional forms. The situation is even more complex with adjectives which – in addition to case, number and gender – can also express degree and definiteness. One single Slovene adjective *pameten* can therefore show no less than 164 different inflected forms where English, for instance, would only have three: "wise", "wiser", "wisest".

With the abundance of different inflected forms it is predictable that the language would not be strict in fixing the word order in sentences. As in most of Slavic languages, sentence elements can be permuted and found in almost all positions. However, different possibilities usually imply that different elements will be emphasised in the sentence, a phenomenon sometimes called topicalisation. A simple five-word sentence *Eva je Adamu dala jabolko* [Eve gave an apple to Adam], composed of a subject, a direct and an indirect object, and a predicator with an auxiliary verb plus a participle forming past tense, can thus produce no less

than 120 permutations, some of which are used to make questions, some sound rather odd, some would imply poetic use, but almost all are legitimate in a specific context. All language technology applications for Slovene are affected by these features, particularly by complex morphology and the free word order implying topicalisation issues, together with the rather complex relation between written and spoken language.

2.2 Slovenian in the Digital Sphere

In 2021, 93% of households in Slovenia had access to the Internet. Statistics also shows that 85% of people aged between 16 and 74 use the Internet every day or almost every day, and 74% of them ordered or bought goods or services online in the last 12 months. In Q1 2021, 69% of people aged between 16 and 74 year responded they used a governmental websites or mobile apps in the last 12 months, and 38% had completed and submitted at least one official electronic form via this method, 87% more than in the same period in 2019, before the COVID-19 pandemic (Statistical Office of the Republic of Slovenia, 2021).

Register.si,[3] the registry for top-level Slovene domain names (.si), operating within Academic and Research Network of Slovenia, reported that in 2020 the overall number of .si domains was 140,702, with an extraordinary increase of 10% in the same year, probably due to the COVID-19 pandemic. The registry reported that 72% of domain holders are companies, the rest are owned by physical persons. The .si domains account for 49.9% of all domains registered in Slovenia, therefore it represents the first choice for companies operating in Slovenia.

Slovene is used extensively on the web and in social media, partly due to the requirement in the Public Use of the Slovene Language Act that the content on web pages administered by Slovene network operators must not be presented and advertised only in foreign languages.[4] Such computer-mediated communication in Slovene differs from standard language variant – as was shown in extensive studies,[5] this representing an additional challenge to the language technologies.

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 AI*.

[3] https://www.register.si/en/
[4] Public Use of the Slovene Language Act (http://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO3924)
[5] Linguistic Analysis of Nonstandard Slovene (https://nl.ijs.si/janes/english/)
[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).

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

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

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

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

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

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

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

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[7]

4 Language Technology for Slovene

In general, it can be acknowledged that since the publication of the White Paper "Slovene Language in the Digital Age" some ten years ago (Krek, 2012) which stated that "in the last two decades language technology for Slovene was never supported by a consistently devised national funding scheme", in the intermittent period the government in Slovenia recognised the need to support LT for Slovene, and language technology did enter into national planning documents and funding schemes (cf. Section 4.3). On the other hand, the number of private companies in Slovenia specialising in LT for Slovene remains extremely low, and most of the LT products described below come either from the (Slovene) academic sphere with national or EU funding, or from the big international IT companies that cover a large number of languages.

4.1 Language Data

Monolingual Corpora

A useful place to consult or download Slovene corpora are the CLARIN.SI NoSketch Engine[8] and KonText[9] concordancers maintained by the CLARIN.SI infrastructure[10] and its repos-

[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).

[8] https://clarin.si/noske/

[9] https://clarin.si/kontext/corpora/corplist

[10] https://www.clarin.si/info/about/

itory.[11] At the time of writing, there are 76 corpora of varying sizes containing Slovene data in the repository, and 59 corpora in the concordancers. Most of them are available for download under open licenses. The more important families of corpora cover general written standard language (Gigafida), Slovene Web and social media (slWaC, Janes), academic discourse (KAS), parliamentary transcriptions (siParl, ParlaMint), Slovene Wikipedia (CLASSLAWiki-sl), historical texts (IMP), literature (MAKS, ELTeC-slv), specialised domains (KoRP, DSI, Konji, etc.), school essays (Šolar, SBSJ). Some of the biggest corpora are now combined in the unified 3.6 billion-word *metaFida* corpus.[12] There are also various manually annotated training and evaluation corpora available (ssj500k, etc.).

Standard written language: A line of corpora of general written Slovene (FIDA, FidaPLUS, Gigafida) was first funded by private initiative in 1997-2000, by a series of national research projects in 2003-2006, and by an extensive "Communication of Slovene" LT project in 2008-2013.[13] The FIDA corpus line grew from 100 million words in 2000 to a more than 1 billion word corpus in 2018. Gigafida is maintained by the Centre of Language Resources and Technologies at the University of Ljubljana and is now updated regularly.[14] It is available in several places, in addition to NoSketch Engine and KonText also in a native specialised concordancer.[15] As a linguistically processed data set Gigafida is available to those that sign a special contract protecting copyright of the original text owners. A 100-million word subset of the corpus (ccGigafida) is available for download in the CLARIN.SI repository under CC BY-NC-SA 4.0 Creative Commons license.[16]

Web and social media: slWaC is a web corpus collected from the .si top-level domain. The current version from 2014 contains 1.2 billion tokens and is annotated with the lemma and the morphosyntax layer. The corpus is distributed under the CC-BY-SA license, and can be obtained by contacting the author.[17] Janes corpus family of texts from Slovene web sites and social media contains blogs, forums, news comments, Wikipedia comments and tweets, altogether around 250 million words.[18] Corpora are available under CC BY 4.0 licence from the CLARIN.SI repository, except for Twitter corpus which is distributed in an encoded version due to Twitter terms-of-service.

Academic discourse: The KAS corpus (1.33 billion words) of Slovene academic writing consists of 65,000 BSc/BA, 16,000 MSc/MA and 1,600 PhD theses written between 2000-2018 and gathered from the digital libraries of Slovene higher education institutions via the Slovene Open Science portal.[19] The corpus is available for download in the CLARIN.SI repository under CLARIN ID-BY-NC-INF-NORED license which requires authentication of the user, acknowledgement of authorship, non-commercial use, informing the rights holder about the intended use, and redistribution is not permitted.[20]

Parliamentary transcriptions: The siParl corpus contains minutes of the Assembly of the Republic of Slovenia from 1990 to 2018, and minutes of the Council of the President of the National Assembly from 1996-2018. The corpus contains 200 million words and meta-data about the speakers, a typology of sessions etc. and structural, editorial and linguistic annota-

[11] Slovene corpora in CLARIN.SI repository (filtered): https://www.clarin.si/repository/xmlui/discover?filtertype_0=language&filter_relational_operator_0=equals&filter_0=Slovenian&filtertype=type&filter_relational_operator=equals&filter=corpus
[12] https://www.clarin.si/kontext/query?corpname=mfida01
[13] http://eng.slovenscina.eu
[14] https://www.cjvt.si/en/research/cjvt-projects/gigafida-corpus/
[15] https://viri.cjvt.si/gigafida/
[16] http://hdl.handle.net/11356/1035
[17] http://nlp.ffzg.hr/resources/corpora/slwac/
[18] Janes blog: http://hdl.handle.net/11356/1138, Janes forum: http://hdl.handle.net/11356/1139, Janes news: http://hdl.handle.net/11356/1140, Janes wiki: http://hdl.handle.net/11356/1137, Janes tweet: http://hdl.handle.net/11356/1142
[19] http://openscience.si
[20] http://hdl.handle.net/11356/1244

tions. ParlaMint 2.1 is a multilingual set of 17 comparable corpora containing parliamentary debates mostly starting in 2015 and extending to mid-2020, with each corpus being about 20 million words in size. Both siParl[21] and ParlaMint[22] are available under CC BY 4.0 license and downloadable from the CLARIN.SI repository.

Historical texts: The IMP digital library contains historical Slovene books and other publications, together 658 texts with over 45,000 pages from the period 1584-1919. These texts were annotated to be used as a language corpus containing 17.7 million tokens. Each word is marked-up with its modernised form, lemma, and morphosyntactic description. The corpus can be downloaded from the CLARIN.SI repository.[23]

Other: Various other types of corpora, e. g. Wikipedia corpus, corpora with literary texts, specialised corpora, corpora with school essays, and various manually annotated training and evaluation corpora can be found in the CLARIN.SI repository under the Creative Commons license.

As the repository is an established place to deposit corpora in the Slovene LT community, one can expect that by querying or harvesting the repository most of the Slovene corpora that are available for download would be found there also in the future.

Multimodal Corpora (audio, video)

The GOS (GOvorjena Slovenščina – Eng. Spoken Slovene) family of corpora contains transcriptions of spoken Slovene. The original GOS includes the transcripts of approximately 120 hours of speech recorded in various situations: radio and TV shows, school lessons and lectures, private conversations between friends or within the family, work meetings, consultations, conversations in buying and selling situations, etc. All speech is transcribed in two versions – with pronunciation-based spelling and with standardised spelling – and it comprises over one million words. Transcriptions are available under CC BY-NC-SA 4.0 license. Recordings are available to those that sign a special contract protecting personal information of the speakers.[24]

GOS VideoLectures is an add-on to the GOS corpus and covers public academic speech (55 lectures and 22 hours). Both transcriptions and recordings available under Creative Commons license and downloadable from the CLARIN.SI repository, similar to other resources that include Dialogue act annotated spoken corpus GORDAN (1 hour), Speech Database of Spoken Flight Information Enquiries SOFES (appr. 10 hours) and SNABI database for continuous speech recognition.[25] SI TEDx-UM is a spoken language resource built from TEDx Talks. The speech database contains 242 talks in total duration of 54 hours. Transcriptions and recordings are available under Creative Commons license on the The Institute of Electronics and Telecommunications, University of Maribor web pages.[26] Video data in CLARIN.SI repository is included in the Multimodal corpus EVA consisting of one episode of an audio/video session plus corresponding orthographic transcriptions with a duration of 57 minutes.[27] There is also some Slovene speech data that can be bought from ELRA, e. g. BNSI Broadcast News (36 hours) consisting of TV news shows from the archive of a Slovene national broadcaster RTV Slovenia.[28]

[21] http://hdl.handle.net/11356/1300
[22] http://hdl.handle.net/11356/1431
[23] http://hdl.handle.net/11356/1031
[24] Spoken corpus GOS: http://hdl.handle.net/11356/1438
[25] CLARIN.SI repository filtered for audio data: https://www.clarin.si/repository/xmlui/discover?filtertype=type&filter_relational_operator=equals&filter=audio
[26] TED Talks: SI TEDx-UM: https://ietk.feri.um.si/en/portfolio/sitedxumenglish/
[27] Multimodal corpus EVA: http://hdl.handle.net/11356/1311
[28] BNSI Broadcast News: http://catalog.elra.info/en-us/repository/browse/ELRA-S0275/

Bilingual/Parallel data

In terms of the availability of parallel data, Slovene has benefited from its status of one of the official EU languages since 2004 and is included in the standard multilingual parallel data sets produced either by EU institutions: JRC-Acquis, DGT-Acquis, DCEP, DGT-TM, EAC-TM, ECDC-TM, JRC-Names,[29] or by EU-funded and other projects: INTERA, WIT3, ParaCrawl, CommonCrawl, OpenSubtitles etc. which are available either from the OPUS web site,[30] or from other data repositories such as ELG[31] or HuggingFace.[32] CLARIN.SI repository offers some early results of the ongoing Development of Slovene in Digital Environment project (cf. Section 4.3), the slenWaC 1.0 corpus of parallel Slovene-English texts crawled from the .si top-level domain, and a Slovene-English corpus for the evaluation of machine translation.[33] Two TM corpora produced by the Secretariat-General of the Slovene government were made available in the context of the ELRC project and are uploaded in the ELRC-Share repository.[34]

Lexical/conceptual resources

There are 82 lexical/conceptual resources with Slovene data in the CLARIN.SI repository available under open access licenses.[35] The ones that deserve special mention due to their size or importance are: *Sloleks* – morphological lexicon containing approx. 100,000 most frequent Slovene lemmas, their inflected or derivative word forms (2.7M) and the corresponding grammatical description. Sloleks 2.0 includes accents automatically assigned by the use of neural networks and partially manually corrected, as well as automatically generated IPA (International Phonetic Alphabet) and SAMPA (Speech Assessment Methods Phonetic Alphabet) transcriptions on lemmas and word-forms. *sloWNet* is the Slovene WordNet developed in the expand approach: it contains the complete Princeton WordNet 3.0 and over 70,000 Slovene literals. These literals have been added automatically using different types of existing resources, such as bilingual dictionaries, parallel corpora and Wikipedia. 33,000 literals have been subsequently hand-validated. *Dictionary of the Slovenian Normative Guide* is a normative orthographic dictionary of Slovene standard language. In 92,617 entries it contains 140,266 lemmas and sublemmas. The entries contain information on spelling, pronunciation, inflection, part of speech, normative information, synonyms, valency. Some semantic indicators, labels and usage examples are also provided. *Thesaurus of Modern Slovene* is an automatically created thesaurus from Slovene data available in a comprehensive English–Slovene dictionary, a monolingual dictionary, and a corpus. It contains 105,473 entries and 368,117 synonym pairs. Other lexical resources include various dictionaries (collocations, terminology, bilingual dictionaries, smaller monolingual dictionaries of general language), frequency lists, etc. Open access data provided by the Centre for Language Resources and Technologies (Sloleks, Thesaurus, etc.) can be browsed on the dictionary portal,[36] and there are two portals with general and terminological dictionaries where both open and propri-

[29] EU Language Technology Resources: https://ec.europa.eu/jrc/en/language-technologies/
[30] https://opus.nlpl.eu
[31] European Language Grid repository (filtered for Slovene parallel corpora): https://live.european-language-grid.eu/catalogue/?resource_type__term=Corpus&language__term=Slovenian&intended_application__term=Machine%20Translation
[32] HuggingFace Datasets (filtered for Slovene parallel corpora): https://huggingface.co/datasets?languages=languages:sl&task_ids=task_ids:machine-translation&sort=downloads
[33] CLARIN.SI, Parallel corpus EN-SL RSDO4: http://hdl.handle.net/11356/1457, slenWaC 1.0: http://hdl.handle.net/11356/1061
[34] https://elrc-share.eu
[35] CLARIN.SI – Lexical Conceptual Resources (filtered for Slovene): https://www.clarin.si/repository/xmlui/discover?&filtertype_0=type&filtertype_1=language&filter_relational_operator_1=equals&filter_relational_operator_0=equals&filter_1=Slovenian&filter_0=lexicalConceptualResource&rpp=100&sort_by=dc.date.issued_dt&order=asc
[36] https://viri.cjvt.si/

etary and copyrighted lexical data can be browsed: Fran, maintained by the Institute of the Slovenian language Fran Ramovš, and Termania, maintained by the Amebis company.[37]

Models and grammars

The first language model available for Slovene was fastText induced from a large collection of Slovene corpora: Gigafida, Janes, KAS, slWaC etc. The fastText embeddings are based on the skip-gram model trained on 3.5 billion tokens of running text.[38] Chronologically, this model was followed by ELMo trained on the Gigafida corpus for 10 epochs. The model can also infer out-of-vocabulary words, since the neural network input is on the character level.[39] The most recent one is the Slovene RoBERTa (A Robustly Optimized Bidirectional Encoder Representations from Transformers) model. The corpora used for training the model have 3.47 billion tokens in total. The subword vocabulary contains 32,000 tokens.[40] Multilingual models are also available, e. g. trilingual BERT (Bidirectional Encoder Representations from Transformers) model, trained on Croatian, Slovene, and English data.[41]

4.2 Language Technologies and Tools

Text Analysis

The standard and most accurate text processing tool for Slovene is the CLASSLA fork of Stanza linguistic annotation pipeline.[42] The pipeline supports processing of both standard and non-standard Slovene on the level of tokenisation and sentence segmentation, part-of-speech tagging, lemmatisation, dependency parsing and named entity recognition. It is available on GitHub and via pip, the Python package manager. Its reported accuracy is 97.39% for part-of-speech (POS) tagging, which includes the complex Slovene morphological features (cf. Section 2.1), 99.1% for lemmatisation, and 92.09% for dependency parsing (Universal Dependencies).

Other tools for processing Slovene include semantic role labeling software, a tool for text normalisation via character-level machine translation, for predicting the level of linguistic and technical standardness, diacritic restoration tool, truecaser for computer-mediated-communication (and other types of) data, etc. All of them are available in the CLARIN.SI GitHub repository.[43]

Spelling checkers for Slovene are available in the popular text processing tools: Microsoft Office, Google Docs, OpenOffice, etc. There is a commercial (spelling and) grammar checker available from the Amebis company, and Slovene is partially supported by the LanguageTool offered in a freemium model.[44] Slovene has a complex system of rules for comma placement. Recently, an open source tool was published by the Centre for Language Resources and Technologies (University of Ljubljana) that works with 94% accuracy based on BERT language model and adapted for comma placement verification problem.[45]

[37] Dictionary portals: Fran https://www.fran.si, Termania https://www.termania.net
[38] fastText: http://hdl.handle.net/11356/1204
[39] ELMo: http://hdl.handle.net/11356/1257
[40] RoBERTa: http://hdl.handle.net/11356/1397
[41] Trilingual BERT: http://hdl.handle.net/11356/1330
[42] CLASSLA is available from https://github.com/clarinsi/classla and the Python package manager from https://pypi.org/project/classla/
[43] CLARIN.SI GitHub repository: https://github.com/clarinsi/
[44] Spelling and grammar checkers: Amebis – BesAna: https://www.amebis.si/besana, LanguageTool: https://languagetool.org
[45] Automated comma placement tool Vejice: https://orodja.cjvt.si/vejice/home, https://github.com/clarinsi/vejice

Speech Processing

There are some Slovene LT companies that develop speech-to-text and text-to-speech tools.[46] Also, Slovene is available in speech technology services offered by big IT companies such as Microsoft and Google,[47] also by some other companies specialising in speech technology.[48] These solutions also found their way to some specialised devices covering many languages.[49] At the University of Ljubljana, a system is developed for automatically translating lectures from Slovene to other languages in real time, in the context of the Online Notes project.[50]

Translation Technologies

Similar to speech, machine translation services for Slovene are available through more or less the same stakeholders – some Slovene LT companies,[51] the big IT companies such as Microsoft and Google,[52] and some other international companies specialising in machine translation technology or general translation services.[53] As an official EU language, Slovene is included in the eTranslation service offered by the European Commission. It is also part of the previously mentioned Online Notes system.

Information Extraction, Language Generation, Human-Computer Interaction

As opposed to translation and speech technologies, Slovene is not yet included in virtual assistant services of big IT companies such as Google Assistant, Microsoft Cortana, Apple Siri etc.[54] Nevertheless, the most active field of the three is the Human-Computer Interaction, with several virtual assistents, chatbots and conversational agents developed in recent years. Most of them were created as part of research project and remain in the experimental stage.[55] There is one chatbot solution for Slovene developed by the Amebis company.[56] Summarisation, or information extraction systems developed specifically for Slovene are in development in the project Development of Slovene in Digital Environment and will be available in 2023 (cf. Section 4.3).

4.3 Projects, Initiatives, Stakeholders

Historically, widespread language technology applications began with the first Slovene spelling checkers at the beginning of 1990s. The first international and national funding came a few years later with the participation in Multext-East project which provided the first Slovene

[46] Amebis, Alpineon – eBralec: https://ebralec.si, Vitasis – Truebar: https://vitasis.si/products/truebar
[47] https://cloud.google.com/speech-to-text/docs/languages, https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/language-support
[48] NEWTON Technologies: https://www.newtontech.net/en/languages/, Sonix: https://sonix.ai/languages/transcribe-slovenian-audio
[49] Pocketalk: https://europe.pocketalk.com/languages-countries/
[50] University of Ljubljana – Online Notes: https://www.cjvt.si/en/infrastructure-support/tolmac/
[51] Vitasis – Truebar: https://vitasis.si/products/truebar, Aikwit: https://aikwit.com/about-us/, Taia: https://taia.io/machine-translation/
[52] Google Cloud Translation support: https://cloud.google.com/translate/docs/languages, Microsoft Azure Translation service: https://docs.microsoft.com/en-us/azure/cognitive-services/translator/language-support
[53] DeepL Translate: https://www.deepl.com/en/translator, Pangeanic: https://pangeanic.com/languages/slovenian-translation-services/, etc.
[54] Google Assistant languages: https://developers.google.com/assistant/console/languages-locales, Microsoft Cortana languages: https://support.microsoft.com/en-us/topic/cortana-s-regions-and-languages-ad09a301-ce3a-aee4-6364-0f0f0c2ca888, Apple Siri languages: https://www.apple.com/ios/feature-availability/#siri
[55] EVA conversational agent: https://dsplab.feri.um.si/en/development-of-personalized-conversational-agents/, PERSIST chatbot as a microservice: https://dsplab.feri.um.si/en/ai-based-chatbot-systems/
[56] Amebis SecondEgo chatbot: https://www.amebis.si/secondego

language resources with standard markup and annotation, and these were later expanded and upgraded in the ELAN (European Language Activity Network), TELRI I in II (Trans European Language Resources Infrastructure) and Concede (Consortium for Central European Dictionary Encoding) projects. At the same time, speech technologies began to be funded in international (SQEL: Spoken Queries in European Languages 1995-1997) and national projects (ARTES, ARGOS, etc.). This trend continued in 2000s, when the Alpineon software company led a consortium in the VoiceTran project (2004-2008) Žganec Gros et al. (2005). In the same period, University of Maribor participated in the EU LC-Star project (Lexica and Corpora for Speech-to-Speech Translation Components), as well as some other EU projects.

With the EU membership of Slovenia in 2004 funding opportunities increased, and Slovene government decided to fund an LT-oriented project (Communication in Slovene, 2008-2013) which produced many of the tools and resources described above. A major breakthrough was achieved when Slovene CLARIN consortium received stable funding in 2015, and entered into the European CLARIN ERIC infrastructure. All major Slovene institutions involved in the development of LT resources, tools and services are members of the CLARIN.SI consortium. The main services provided by CLARIN.SI infrastructure are its CTS (Core Trust Seal) and CLARIN certified repository for language resources and tools, the NoSketch Engine and KonText concordancers, as well as services for automated text annotation for Slovene and other South Slavic languages, services for manual text annotation (WebAnno), for storage, download and cooperative development of language resources and technologies (the CLARIN.SI GitLab installation, the CLARIN.SI virtual organisation on GitHub), and for knowledge transfer, primarily the CLARIN CLASSLA Knowledge centre jointly run by CLARIN.SI and the Bulgarian CLARIN.

From 2018, Slovene Ministry of Culture started financing several projects which produced different language resources now available in CLARIN.SI repository, among them school web portals Franček and Slovenščina na dlani.[57] However, the biggest investment in LT for Slovene is the "Development of Slovene in Digital Environment" project financed by the Slovene Ministry of Culture between 2020-2023.[58] The project will significantly upgrade existing LT resources, tools and services, or produce those that do not exist yet. Results include:

- *text processing resources and tools*: significant upgrades of the majority of existing corpora and text analysis tools described above, availability of the SuperGLUE benchmark for Slovene, word embeddings model(s);
- *speech processing resources and tools*: 1000 hours of transcribed speech, ASR models for general language and two specialised applications, online ASR service, syntactic and acustic normaliser, grapheme to phoneme transformer, punctuator;
- *semantic resources and tools*: digital dictionary database (with semantic, morphological, phonetic and other lexical data, available through a REST API), tools for named entity recognition, coreference resolution, relation extraction, word sense disambiguation, semantic shift detection, automatic summarisation and question-answering;
- *machine translation*: increased availability of parallel (translation memory) corpora, open access neural machine translation models, online machine translation service, anonymisation tools;
- *terminology management*: national terminology portal with an online editor, a terminology extraction tool and a concordancer for specialised corpora, open access terminology data, REST API service;

[57] Franček: https://www.xn--franek-l2a.si, Slovenščina na dlani: https://slo-na-dlani.si/prijava
[58] Razvoj slovenščine v digitalnem okolju (RSDO): https://www.slovenscina.eu

The results of the project are expected to be published on the CLARIN.SI and GitHub repositories in November 2022 and February 2023. Language resources will be available under Creative Commons open access licenses, tools will be available under Apache open source licenses. An important result of the project is a long-term plan for LT support for Slovene. It is expected that this will be reflected in the major governmental planning documents for the current period.[59]

5 Cross-Language Comparison

The LT field[60] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded results unforeseeable before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[61] 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.)

[59] National programme on encouraging the development and use of artificial intelligence by 2025: http://www.ds-rs.si/sites/default/files/dokumenti/npai_si_2021-03-10_cistopis_zdsma.pdf, Resolution on the National Programme for Language Policy 2021-2025: http://www.pisrs.si/Pis.web/pregledPredpisa?id=RESO123, etc.

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

[61] 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.

5.2 Levels of Technology Support

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

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[62]

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[63] 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.[64]

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

[62] 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.

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

[64] 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.

level of technology support is subject to change and it reflects a snapshot of the available resources on January 2022.

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

5.4 Results and Findings

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

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g. German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[65] 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 2 visualises our findings.

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

The results of the present comparative evaluation reflect, in terms of distribution and imbalance, the results of the META-NET White Paper Series (Rehm and Uszkoreit, 2012). The complexities of the analyses clearly differ across 2012 and 2022 studies, and as such, a direct comparison between the two studies can therefore not be made. However, we can instead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e. g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

The results of this analysis are only informative of the relative positioning of languages, but not of the progress achieved within a specific language. The LT field as a whole has

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

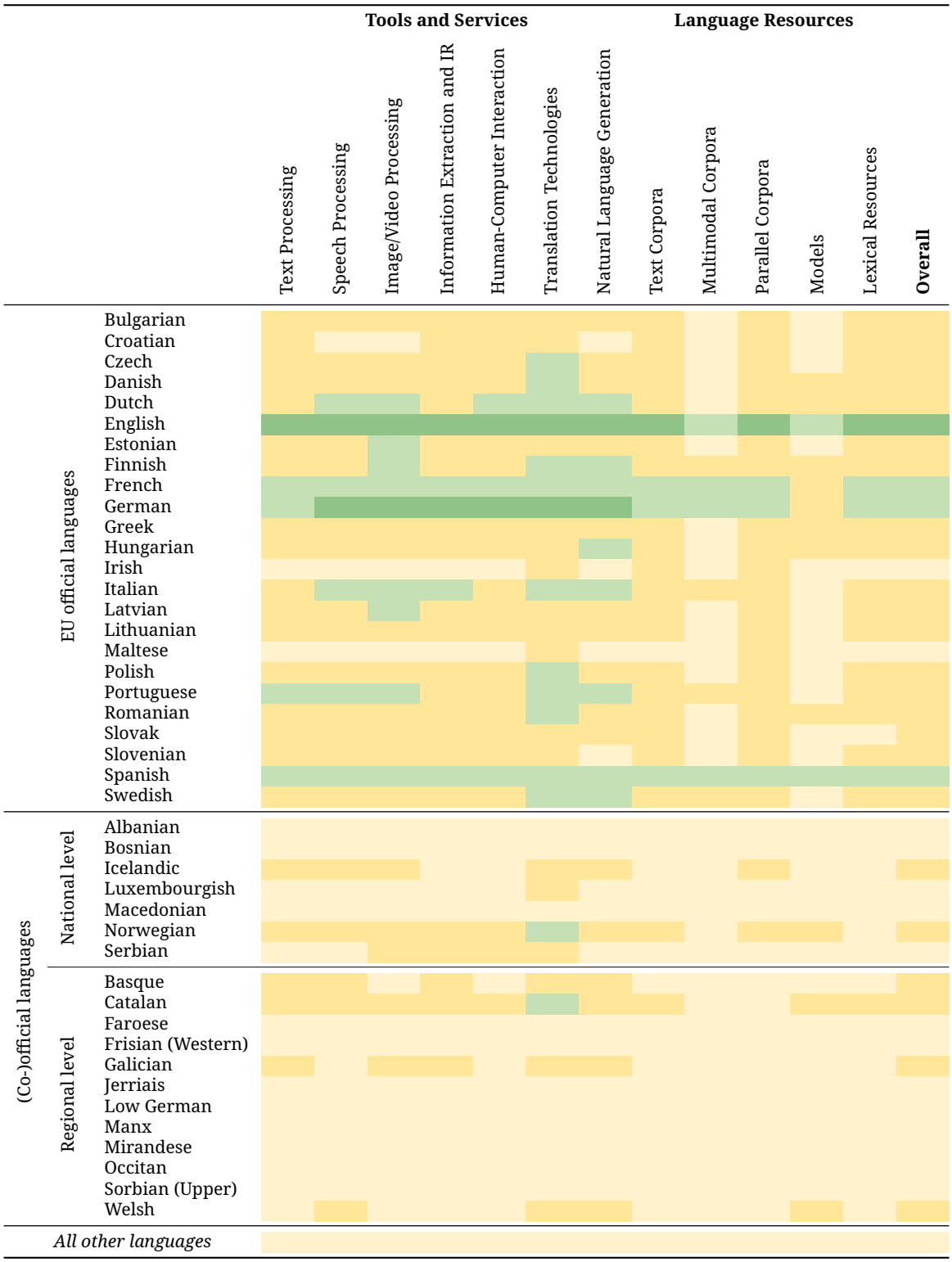

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

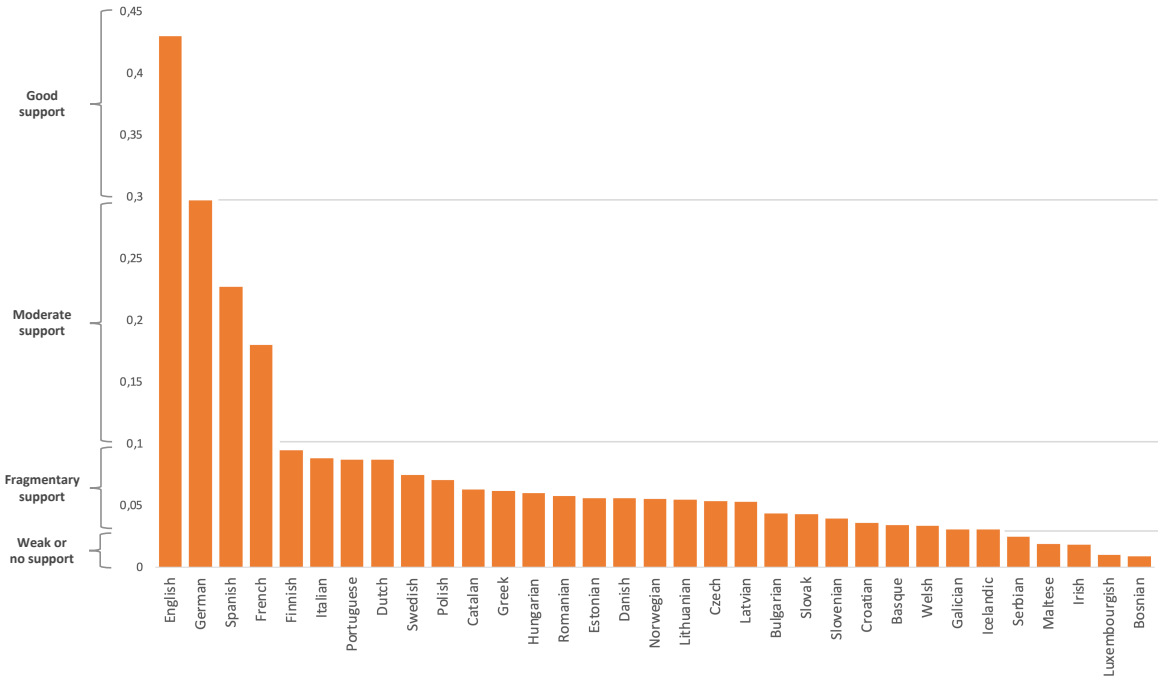

Figure 2: Overall state of technology support for selected European languages (2022)

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

6 Summary and Conclusions

The analysis of available resources and the cross-language comparison show that Slovene has fragmentary LT support, along with other official EU languages with a smaller number of speakers, such as Croatian, Slovak, Bulgarian, Latvian, etc. In particular, missing types of language resources that the analysis brought to the fore and need attention in the future are (language) models and multimodal corpora. In relation to tools and services, the underdeveloped area is Natural Language Generation (summarisation, paraphrasing, text re-writing, simplification, generation of questions etc). It can be expected that the ongoing project – Development of Slovene in a Digital Environment – will to some extent cover the creation of language models and NLG tools, in particular summarisation. However, multimodal corpora and other types of NLG tools are not covered in current plans.

In the overall state of technology support for selected European languages chart (Figure 2) Slovene is positioned rather low, in 23rd place, close to the better supported non-official EU languages such as Basque or Galician, well behind the comparable official EU languages such as Estonian, Latvian and Lithuanian, and still further from languages like Swedish, Greek or Hungarian which have approx. five times the number of speakers as Slovene. The figure clearly shows that Slovenia as an EU state will need to encourage the creation of resources, tools and services, and provide a stable environment for LT development. Currently, most

of the long-term activity is financed through the CLARIN.SI infrastructure which is a major achievement compared with the situation from ten years ago, but still not enough if the current trends in AI are taken into account, with Natural Language Understanding as the goal.

There are several state planning documents that could lead to better support for LT for Slovene in the future. The first is the *Resolution on the National Programme for Language Policy 2021-2025* which includes an extensive section on language infrastructure covering topics such as language description (dictionaries, lexical resources), standardisation, terminology, multilingualism, language technologies, digitisation, special needs. The total sum of funds assigned to these topics is approximately 15M€. Another important document covering language technology is the *National programme on encouraging the development and use of artificial intelligence by 2025*. Language technology is mentioned in several sections and there are funds assigned to topics covering LT. However, it is not clear what portion of funding will be assigned to LT in a narrower sense, or which mechanisms will be used.

In general, one can conclude that (a) the support for Slovene is comparable with other languages with a similar status, (b) there is a general awareness in the governmental bodies that LT for Slovene should be supported in the future, (c) the LT community is growing, also through new educational initiatives such as the MA study of Digital Linguistics (Faculty of Arts, University of Ljubljana), (d) there is an infrastructural support, mainly through CLARIN.SI infrastructure at the Jožef Stefan Institute, which also covers all other stakeholders through CLARIN.SI consortium. However, more efforts are needed in the future to bring the existing support closer to the one available for other (official EU) languages.

References

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

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

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

Simon Krek. *Slovenski jezik v digitalni dobi – The Slovene Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London, 9 2012. URL http://www.meta-net.eu/whitepapers/volumes/slovene. Georg Rehm and Hans Uszkoreit (series editors).

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.

SURS Statistical Office of the Republic of Slovenia. (Religious, linguistic and national structure of the population of Slovenia - Censuses 1921-2002), 2003. http://www.stat.si/popis2002/gradivo/2-169.pdf.

SURS Statistical Office of the Republic of Slovenia. (Usage of internet in households and by individuals, detailed data, 2021), 2021. https://www.stat.si/StatWeb/en/News/Index/9892.

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.

Jerneja Žganec Gros, France Mihelič, Tomaž Erjavec, and Špela Vintar. The voicetran speech-to-speech communicator. In *TSD*, 2005.