EUROPEAN LANGUAGE EQUALITY

D1.23

Report on the Lithuanian Language

Authors	Anželika Gaidienė, Aurelija Tamulionienė
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	D1.23
Deliverable title	Report on the Lithuanian 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
Authors	Anželika Gaidienė, Aurelija Tamulionienė
Reviewers	Jaroslava Hlavacova, Maria Giagkou
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 **2**

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

3 What is Language Technology? **5**

4 Language Technology for Lithuanian **7**
4.1 Language Data 7
4.2 Language Technologies and Tools 9
4.3 Projects, Initiatives and Stakeholders 10

5 Cross-Language Comparison **13**
5.1 Dimensions and Types of Resources 13
5.2 Levels of Technology Support 14
5.3 European Language Grid as Ground Truth 14
5.4 Results and Findings 15

6 Summary and Conclusions **17**

List of Figures

1 Overall state of technology support for selected European languages (2022) . . 17

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) 16

List of Acronyms

AI	Artificial Intelligence
CEF	Connecting Europe Facility
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
DL	Deep Learning
DH	Digital Humanities
EC	European Commission
EFNIL	European Federation of National Institutes for Language
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)
ELRC	European Language Resource Coordination
HPC	High-Performance Computing
LR	Language Resources/Resources
LSSAIS	Lithuanian Syntactic and Semantic Analysis Information System
LT	Language Technology/Technologies
ML	Machine Learning
MT	Machine Translation
NLP	Natural Language Processing
NLU	Natural Language Understanding
SR	Speaker Recognition

Abstract

In recent years, the knowledge society has been entering a qualitatively new phase, one marked by the rapid development of advanced information technologies, in particular the collection and processing of big data and the development of technologies based on artificial intelligence. Information technology is being increasingly used in all major areas of the society, such as public administration and the judiciary, education, science, culture and heritage, the media, e-banking, health, energy, public transport, nature protection, national defence, business, etc. At the same time, these changes pose new challenges, the most important of which are to strengthen synergies, ensure cyber security and protection against the spread of disinformation, develop distance learning, improve the quality of life, reduce language barriers, address employability and aging, and reduce social and regional disparities. Language technology is an important part of information technology and one of the necessary tools to address these challenges (State Commission of the Lithuanian Language, 2020).

The volume of linguistic communication between people and computers is growing at such a pace, that it is necessary to ensure the full existence of the Lithuanian language as a means of communication in the digital world.

On 13 October 2020, the Seimas of the Republic of Lithuania approved an important document for the Lithuanian language and its future: *The Guidelines for the Development of the Lithuanian Language in the Digital Environment and the Progress of Language Technologies for 2021 – 2027*. The Guidelines were drafted by a working group formed by the State Commission of the Lithuanian Language. These Guidelines will help to ensure the full use of the Lithuanian language in the digital environment and to establish and maintain the status of the Lithuanian language in the information society. This requires additional digital language resources – texts and speech corpora, the development of language technologies and their integration in public services based on them, so that no group in the society or region experiences the digital divide, and also that foreign languages can integrate more easily into Lithuanian society. Language technologies must help strengthen the ties between Lithuanian society and the diaspora and reduce the exclusion of the Lithuanian-speaking community in the global knowledge society (Jaroslavienė and Miliūnaitė, 2020).

Significant progress has been made in adapting the Lithuanian language to the digital environment: a number of digital language resources and basic language analysis tools, as well as complex online language services and the Lithuanian language ontology have been developed, while a number of computer programs and tools have been localised. Computer applications relevant to the society is being Lithuanianised, and the standardisation of computer terms is being carried out. Lithuanian researchers actively participate in the cooperation and mobility activities of international associations, and a core of Lithuanian language specialists working in the field of IT application, and systematically developing innovative works in this field, has been formed. Lithuania also strives for all citizens to have full access to digital solutions, which adds importance to the policy of adapting them for those living with disabilities (State Commission of the Lithuanian Language, 2020).

Anotacija

Pastaraisiais metais žinių visuomenė pereina į kokybiškai naują etapą, kurį žymi sparti pažangių informacinių technologijų plėtra, pirmiausia didžiųjų duomenų kaupimas ir apdorojimas bei dirbtiniu intelektu grįstų technologijų kūrimas. Informacinės technologijos vis plačiau diegiamos visose pagrindinėse visuomenės veiklos srityse, tokiose kaip valstybės administravimas ir teismų sistema, švietimas, mokslas, kultūra ir jos paveldo saugojimas, žiniasklaida, elektroninė bankininkystė, sveikatos apsauga, energetika, viešasis transportas,

gamtosauga, krašto apsauga, verslas ir kt. Kartu šie pokyčiai kelia ir naujų uždavinių, iš kurių svarbiausi – glaudinti šių sričių sąveiką, užtikrinti kibernetinį saugumą ir apsaugą nuo dezinformacijos sklaidos, plėtoti nuotolinį mokymą, gerinti gyvenimo kokybę, mažinant kalbų barjerus, sprendžiant galėjimo įsidarbinti ir visuomenės senėjimo problemas, mažinant socialinę ir regionų atskirtį. Kalbos technologijos yra svarbi informacinių technologijų dalis ir vienas iš būtinų įrankių šiems uždaviniams spręsti (State Commission of the Lithuanian Language, 2020).

Kalbinės komunikacijos apimtys tarp žmonių ir išmaniųjų įrenginių auga tokiais tempais, kad būtina užtikrinti lietuvių kalbos, kaip tokios komunikacijos priemonės, visavertį gyvavimą skaitmeniniame pasaulyje.

2020 m. spalio 13 d. Lietuvos Respublikos Seimas patvirtino lietuvių kalbai ir jos ateičiai svarbų dokumentą – Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021 – 2027 metų gaires. Jas parengė Valstybinės lietuvių kalbos komisijos sudaryta darbo grupė. Šios gairės turi padėti užtikrinti visavertį lietuvių kalbos vartojimą skaitmeninėje terpėje, įtvirtinti ir palaikyti lietuvių kalbos statusą informacinėje visuomenėje. Tam reikia gausinti skaitmeninius kalbos išteklius – tekstynus ir garsynus, plėtoti kalbos technologijas ir jų pagrindu kurti viešąsias paslaugas, kad nė viena visuomenės grupė ar regionas nejaustų skaitmeninės atskirties, o kitakalbiai galėtų lengviau integruotis į Lietuvos visuomenę. Kalbos technologijos turi padėti stiprinti Lietuvos visuomenės ir išeivijos ryšius, mažinti ir lietuviškai kalbančios bendruomenės atskirtį globalioje žinių visuomenėje (Jaroslavienė and Miliūnaitė, 2020).

Gairėse numatyti esminiai lietuvių kalbos technologijų uždaviniai ar iššūkiai, ką reikėtų artimiausiu metu Lietuvoje daryti, kuriomis kryptimis dirbti:

1. Didinti specialistų, dirbančių kalbos technologijų srityje, kompetenciją ir kelti visuomenės gebėjimo naudotis kalbos technologijų teikiamomis galimybėmis lygį.
2. Kaupti ir gausinti atvirus, patikimus, kokybiškus, pakartotinai pritaikomus skaitmeninius kalbos išteklius ir kitus skaitmeninius kalbos duomenų rinkinius.
3. Plėtoti kalbos technologijų infrastruktūrą, kalbos technologijų taikymą viešajame sektoriuje ir viešosiose paslaugose, kurti ir tobulinti viešai prieinamus informacinių technologijų sprendinius ir priemones.

Šiuo metu pasiekta pastebima lietuvių kalbos pritaikymo skaitmeninei terpei pažanga: parengta nemažai skaitmeninių kalbos išteklių ir pagrindinių kalbos analizės priemonių, sukurta sudėtingų internetinės kalbos paslaugų, sukurta lietuvių kalbos ontologija, lokalizuota nemažai kompiuterinių programų ir įrankių. Lietuvinama visuomenei aktuali taikomoji kompiuterinė programinė įranga, atliekami kompiuterijos terminų norminimo darbai. Lietuvos mokslininkai aktyviai dalyvauja tarptautinių asociacijų bendradarbiavimo ir mobilumo veiklose, sutelktas lietuvių kalbos specialistų, dirbančių informacinių technologijų taikymo srityje ir sistemingai plėtojančių inovatyvius šios srities darbus, branduolys. Lietuva taip pat siekia, kad visi gyventojai turėtų galimybę visavertiškai naudotis skaitmeniniais sprendiniais, todėl itin svarbi jų pritaikymo neįgaliesiems politika (State Commission of the Lithuanian Language, 2020).

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 groundworks for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

More than 40 research partners, who are experts in more than 30 European languages, have conducted an enormous and exhaustive data collection procedure that provides a detailed, empirical and dynamic map of technology support for our languages.[1] The report has been developed in the framework 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 Lithuanian Language in the Digital Age

2.1 General Facts

Lithuanian is a Baltic language from the Indo-European family. Lithuanian and Latvian are the two surviving Baltic languages. Since 2004, Lithuanian has been one of the official languages of the European Union. Lithuanian is the state language of the Republic of Lithuania and is enshrined in the Constitution as such. The use of the state Lithuanian language in public life is regulated by the State Law on the Lithuanian Language (1995). According to this law, the state guarantees education at all levels in the native Lithuanian language. There are also schools for ethnic minority students in Lithuania that teach non-state languages: Russian, Polish, Belarusian, as well as mixed schools.

According to date from 2012, there were about 3.6 million Lithuanian speakers: about 2.7 million in Lithuania, where, in addition to Lithuanians, the language is used by about 350,000 people of other nationalities, and by about 0.6 million people abroad. It is still used by Lithuanian national minorities in the south-east of Latvia, in the north-east of Poland, in the north-west of Belarus, in diaspora communities in Ireland, Australia, Brazil, Estonia, Spain, the United States, Canada, Kazakhstan, Russia, Ukraine and elsewhere.[3] By the number of speakers, the Lithuanian language ranks 144th in the world.

Since 2009, a decrease in the population of Lithuania has been observed, a product of the declining birth rate and emigration.

In 2009 – 2019, 496,300 permanent residents migrated from Lithuania, of whom 453,300 (91.3%) were citizens of the Republic of Lithuania and 43,000 (8.7%) were foreigners. During this period, 225,200 people immigrated to Lithuania, of whom 158,600 (70.4%) were citizens of the Republic of Lithuania and 66,600 (29.6%) were foreigners.[4]

Lithuanian is the most conservative of the Indo-European living languages, and it has best preserved many of its archaic features. From a typological point of view, the Lithuanian language has many unique features, including the abundant forms of variation, the characteristic synthesis of tonal and dynamic stress, and the extremely diverse order of words reflecting the complex syntactic level of discourse communication (Vaišnienė et al., 2012).

The standard Lithuanian language was formed at the beginning of the 20th century on the basis of one of the Aukštaitian dialects. It is characterised by a great variety of regional

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they can be discovered, browsed and further investigated by means of comparative visualisations across languages.
[2] https://european-language-equality.eu
[3] https://www.vle.lt/straipsnis/lietuviu-kalba/
[4] https://osp.stat.gov.lt/statistiniu-rodikliu-analize?hash=18a1efb1-3950-4cff-8112-35d727cbcd3f#/

variants, the two main dialects – Aukštaitian and Samogitian – differing with respect to phonetics, morphology, syntax and vocabulary. These dialects are divided into fourteen larger dialects (varieties) that are further divided into even smaller sub-dialects (Vaišnienė et al., 2012).

The development of Lithuanian spelling began in the 16th century. Based on the prevailing principle of use of letters, the spelling of the Lithuanian language is morphological, but there are also elements of phonological and historical spelling (traditional, mostly vowel spelling). The Lithuanian alphabet was formed in the 16th – 20th centuries on the basis of the Latin alphabet, to which nasal vowels (ą, ę, į, ų) and letters with diacritics (č, š, ž, ė, ū) were added. The current Lithuanian language alphabet has 32 letters: 12 vowels, 20 consonants, and 3 letter combinations (ch, dz, dž). In addition to letters, accents, punctuation, and other signs are used in writing.[5]

The grammatical structure of the Lithuanian language is of a flexural type; the vocabulary is the most variable level of the language. Some words disappear and are replaced by new ones. In the current Lithuanian language, there is a pronounced abundance of terms in various fields. The vocabulary of the Lithuanian language consists of old words, inherited from the Proto-Indo-European language, borrowings, and new words based on inherited words and borrowings.[6]

Among other living Indo-European languages, Lithuanian has the best-preserved synthetic sentence structure. In it, syntactic relations are mainly expressed in interrelated word forms.[7]

At the end of the 19th and the beginning of the 20th centuries, due to its archaic structure and vocabulary, the Lithuanian language became the focus of research of the most famous European scholars of Indo-European languages, mainly linguists from Germany, Poland, Russia and other countries. The Lithuanian language has been taught at universities in Berlin, Jena, Leipzig, Moscow, St. Petersburg, Paris, Prague, Vienna, etc.

At the beginning of the 21st century, there were more than 30 institutions for Lithuanian studies (and sometimes Baltic studies) in various countries (mostly in Europe), where the Lithuanian language (and culture) is being researched and/or taught. At nearly 20 of them, the Lithuanian language (and sometimes literature, folklore or cultural history as well) is taught as an optional subject; sometimes the Lithuanian language as a subject makes a more general part of the philological (historical linguistics, Slavic studies) or other (Eastern European, Northern European and Baltic language and culture) curricula (e. g. at Prague, Pisa, Florence, Berlin Humboldt, Budapest and other universities).[8]

2.2 Lithuanian in the Digital Sphere

According to 2021 data, in the 16 – 74 age group, almost 87% of the Lithuanian population uses the Internet (compared to almost 64.1% in 2011, and 74.4% in 2016); this figure is as high as 100% in the 16 – 24 age group, and 55.2% in the 65 – 74 age group.[9]

According to 2021 data, 81.4% of households have a personal computer, and 86.6% have Internet access.[10]

The Internet is mainly used for information retrieval, communication, entertainment and banking: 79% of the population aged 16 – 74 use the Internet for communication, 74% read the news, 71% use the Internet for entertainment (watch movies or TV shows, listen to music, play or download records, games), 68% use online banking services. 27% of the population

[5] https://www.vle.lt/straipsnis/lietuviu-kalbos-rasyba/
[6] https://www.vle.lt/straipsnis/lietuviu-kalbos-leksika/
[7] https://www.vle.lt/straipsnis/lietuviu-kalbos-sintakse/
[8] https://www.vle.lt/straipsnis/lietuviu-kalbos-studijos-uzsienyje/
[9] https://osp.stat.gov.lt/statistiniu-rodikliu-analize?hash=b3603975-ca07-47cb-aaaf-bc3a3a403a1f#/
[10] https://osp.stat.gov.lt/lietuvos-statistikos-metrastis/lsm-2019/mokslas-ir-technologijos/informacines-technologijos

use the Internet for learning, professional development or self-education purposes.[11] 82.2% of 16 – 74 year olds use the Internet for personal purposes, for example, 65.2% interact with others on social networks, 70.5% of the population correspond in real time (e. g. via Skype, Messenger, WhatsApp, Viber, Snapchat).[12]

In 2021, about 225,000 *.lt* domains were registered (compared to approximately 139,000 in 2012, and approximately 188,100 in 2018), of which more than 2,000 contain distinctive Lithuanian letters (*ė, ž*, etc.).

In addition, Lithuania remains among the leaders in fibre-optic Internet service. In Lithuania, the coverage of the fibre-optic network is 46.8%.[13]

3 What is Language Technology?

Natural language[14] 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.

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 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

[11] https://osp.stat.gov.lt/skaitmenine-ekonomika-ir-visuomene-lietuvoje-2020/gyvenimas-internete
[12] https://osp.stat.gov.lt/statistiniu-rodikliu-analize?hash=1194ea01-82ee-4bde-a222-94c5ced50f4e#/
[13] https://ivpk.lrv.lt/lt/naujienos/lietuva-islieka-tarp-pirmaujanciu-sviesolaidinio-interneto-lyderiu-2
[14] 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 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 a 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.

4 Language Technology for Lithuanian

The level and advancement of language technologies in Lithuania can first and foremost be appraised by the degree of achievement of the goals rooted in the 2014 – 2020 guidelines (State Commission of the Lithuanian Language, 2014) for the expansion of the Lithuanian language in information technologies.[15] Notably, those goals have been achieved with a great deal of success, yet some follow-up actions are needed, depending on the progress of the rapidly shifting language technologies on the global market and amidst the society.

4.1 Language Data

Monolingual / Bilingual Corpora

Lithuania continues to create and develop general resources needed for the purposes of building language technologies and devising their applications. There are a number of **corpora** in Lithuania. The largest corpus of the Lithuanian language is the *Corpus of the Contemporary Lithuanian Language*.[16] It features 140,921,288 words (11.6% of them are from fiction, 14.2% from non-fiction, 10% from administrative literature, 63.8% from publications, 0.3% from spoken language). The corpus was launched in 1992 and was last updated in 2011.

There are also several morphologically and syntactically annotated corpora. The volume of the *Morphologically Annotated Corpus* MATAS[17] is 1.6 million words (36% of them are from publications, 24% from science literature, 19% from fiction, 2.8% from administrative texts, 6.8% from stenographs of the Parliament of the Republic of Lithuania). Entries to the corpus are made in a semi-automated fashion and the outcomes are subject to review by linguists. This corpus has highlighted the immense morphological polysemy of the Lithuanian language, with nearly one half of all forms being morphologically polysemantic. The latest version of the *Syntactically Annotated Treebank* ALKSNIS[18] (ALKSNIS 3.0) consists of 3,643 syntactically annotated sentences in the PML (Prague Mark-up Language) format.

There are a number of parallel corpora developed in Lithuania as well. One *Parallel corpus*[19] consists of Czech – Lithuanian words (20.29%), English – Lithuanian words (76.6%), Lithuanian – Czech words (0.8%), and Lithuanian – English words (2.31%). The *LILA parallel corpus*[20] has been compiled in a semi-automated manner, with texts aligned at a paragraph and sentence level. The corpus features texts published in or after 1991. The total volume of the corpus is 8,782,050 words, its major chunk consisting of texts in the Lithuanian – Latvian languages (3,448,745 words), and texts in the Latvian – Lithuanian languages making up half the number (1,695,160 words). This asymmetrical data structure is the product of the recent predominance of Lithuanian – Latvian rather than Latvian – Lithuanian translations.

Other corpora include *The Corpus of the Spoken Lithuanian Language*;[21] *The compendium of textbook texts* KLASIUS;[22] *The Corpus of Dialects*[23]; *The Corpus of the Old Lithuanian Language*;[24] *The Corpus Academicum Lithuanicum* CorALit, and others.

Most corpora are open access. Nonetheless, considering the demand for language data, it needs to be said that corpus data has to be augmented and new corpora (especially mul-

[15] http://www.vlkk.lt/kalbos-politika/lietuviu-kalbos-pletros-informacinese-technologijose-gaires/lietuviu-kalbos-pletros-informacinese-technologijose-2014-2020-m-gaires
[16] http://tekstynas.vdu.lt/tekstynas/
[17] https://klc.vdu.lt/matas-morfologiskai-anotuotas-tekstynas/
[18] https://klc.vdu.lt/alksnis-sintaksiskai-anotuotas-tekstynas/
[19] https://klc.vdu.lt/lygiagretus-tekstynas
[20] https://klc.vdu.lt/lila-lygiagretusis-tekstynas/
[21] http://sakytinistekstynas.vdu.lt
[22] https://raštija.lt/resurso-katalogas/klasius-v2/
[23] http://tarmiuarchyvas.lki.lt/pradinis.php?sutrump=bnd
[24] http://coralit.lt/node/1

tilingual parallel) should be developed to reflect as many diverse areas of language use as possible.

Lexical Resources

Lithuania continues to develop **digital dictionaries and databanks**. Users have free online access to the latest *The Dictionary of the Standard Lithuanian Language*[25] (74,059 lemmas) as well as other dictionaries, such as *The Dictionary of the Modern Lithuanian Language*[26] (48,342 lemmas); *The Dictionary of the Lithuanian Language*[27] (reflecting the lexis of the Lithuanian language between the 16th century and the late 20th century; a total of 310,659 lemmas); the ongoing *The Database of Lithuanian Neologisms*[28] (featuring new words (loanwords and new coinages), phrases and abbreviations, new meanings of words that have emerged in the Lithuanian language after the end of the 20th century and are currently in public usage, as well as information about the origins, usage, and standardisation of neologisms). Currently, the dictionary has over 7,000 entries. Other digital dictionaries, such as *The Dictionary of Synonyms*,[29] *The Dictionary of Antonyms*,[30] *The Dictionary of Phraseology*,[31] *The Dictionary of Comparisons*,[32] various bilingual (Lithuanian – English; English – Lithuanian; Lithuanian – German; German – Lithuanian, and so on) dictionaries are also freely accessible online. Most of the digital dictionaries are developed by the Institute of the Lithuanian Language and are available in E. KALBA[33] – the information system for Lithuanian language resources (Jaroslavienė and Auksoriūtė, 2019).

Despite this abundance of digital dictionaries, considering the demands of language technologies and of the public, the dictionaries of synonyms, antonyms, and phraseology have to be updated and the dictionaries of pronunciation and combinability, among others, digitalised.

Speaking about **terminological resources**, *The Term Bank of the Republic of Lithuania*[34] – the largest and most reliable, content-wise, source of Lithuanian terms – merits a mention. Its key sources are regulations and glossaries of terms (with more than 255,000 articles on terms available). This bank features sets of terms from 26 different areas such as politics, defence, finance, environment, culture, health, and so on, which are further divided into subcategories. *The Database of Terms of the Lithuanian Standards Board*[35] offers a collection of terms from different standards. Currently, it contains nearly 76,000 articles on terms. RAŠTIJA.LT,[36] the information system for Lithuanian language resources, offers a search functionality across 32 glossaries of terms. The fields of the terms are defined by the entries and range from electrical engineering, computers and management to linguistic didactics, mathematics, metrology, meteorology and others.

Considering the demands of the public, it should be said that the available terminological contents are lacking and additional and up-to-date compendia of terms are needed. The databases of terms vary in their structure and technological solutions, which makes it more difficult to use the data in other technological solutions. Open terminological data is scarce as well.

[25] https://ekalba.lt/bendrines-lietuviu-kalbos-zodynas/
[26] https://ekalba.lt/dabartines-lietuviu-kalbos-zodynas/
[27] https://ekalba.lt/lietuviu-kalbos-zodynas/
[28] https://ekalba.lt/naujazodziai/naujienos
[29] https://ekalba.lt/sinonimu-zodynas/
[30] https://ekalba.lt/antonimu-zodynas/
[31] https://ekalba.lt/frazeologijos-zodynas/
[32] https://ekalba.lt/palyginimu-zodynas/
[33] https://ekalba.lt
[34] http://terminai.vlkk.lt
[35] http://lsd.lt/index.php?-452282422
[36] https://raštija.lt

Semantic networks and ontologies in Lithuania are few. There is the *General Ontology of the Lithuanian Language*, the open-access ontology of Lithuanian medical terms *Snomed CT*,[37] the electronic service *E-terms (Ontologies)*,[38] which includes ontologies in the following fields: *The Ontology of Human Anatomy Terms* (7,500 terms), *The Ontology of Economy Terms* (500 terms), and *The Ontology of Computer Hardware and Parts* (1,000 terms). There are several Lithuanian wordnets that can be developed further, such as *LitWorNet*[39] (for more, see (Vitkutė-Adžgauskienė et al., 2015)) and *WordNet*[40] (with roughly 24,000 synsets, of which 10,000 are linked with the synsets in the *Princeton WordNet*). However, the available ontologies and wordnets are inadequate and have to be enlarged and expanded.

4.2 Language Technologies and Tools

Machine Translation

The ALPMAVIS machine translation system is freely available to the public.[41] In 2018, a new project was launched and has so far seen the development of an open and free translation environment, grounded in neural networks, with improved quality, support for additional pairs of languages, deployment of speech recognition and synthesis solutions. Further actions will include adapting the infrastructure to provide e-government services (State Commission of the Lithuanian Language, 2020). "Tilde Informacinės Technologijos" offers an opportunity to use multilingual machine translation systems based on the latest neural networks, for free.[42] Systems that were built outside of Lithuania should be mentioned, such as *Google Translate*; *Microsoft Bing Translator*, and *eTranslation* (designed for translations in EU languages, which is available for free to governmental institutions and is geared towards translation of administrative and legal texts).

Continued development of machine translation systems would require more bilingual parallel corpora as well as specialised text data to ensure higher quality of machine translation (State Commission of the Lithuanian Language, 2020).

Speech Technology (Speech Recognition and Speech Synthesis)

The available *Lithuanian Speech-to-text Transcription Service*[43] consists of four modules covering different areas: administrative, legal, medical, and standard colloquial. There are also a number of services where speech recognition technology is used to voice-control computers, such as *Browser* (browsing voice control); *Controller* (computer voice control); *Helper* (voice control for the disabled), and so on. The same technology was used to develop the following services: *Coaching Robot Controller*[44] (the controller of a teaching robot for kids); *Caller*[45] (making calls to contacts in a phonebook); *Taxi Caller*[46] (a taxi calling service); *Interlingual Communicator*[47] (a multilingual communicator in Lithuanian – Chinese), and so on.

[37] https://www.snomed.lt/snomed-ct-pritaikomumas-uzsienio-salyse-elektroninis-sveikatos-irasas-ligoninesvaldymo-irankis-palaikomas-snomed-ct/
[38] https://ekalba.lt/esavokos/
[39] http://mackus.vdu.lt/LitWordNet/
[40] https://ekalba.lt/zodziu-prasmiu-tinklas/?p=1
[41] https://vertimas.vu.lt
[42] https://translate.tilde.com/#/
[43] https://semantika.lt/Analysis/Transcriber
[44] https://raštija.lt/liepa-2/paslaugos-vartotojams/ugdanciojo-roboto-valdytuvas/
[45] https://raštija.lt/liepa-2/paslaugos-vartotojams/skambintuvas/
[46] https://raštija.lt/liepa-2/paslaugos-vartotojams/taksi-iskviestuvas/
[47] https://raštija.lt/liepa-2/paslaugos-vartotojams/tarpkalbinis-komunikatorius/

There is also a free speech recognition application available,[48] which converts speech into text using a pre-recorded audio file or real-time dictation.

Some of the services available in Lithuania feature speech synthesis technology; these are *Pronouncer*[49] (an audio dictionary of Lithuanian neologisms); *The Lithuanian Speech Synthesiser for the Blind*[50] (a SAPI5-compatible synthesiser of Lithuanian speech that reads text displayed on a computer monitor); *The Mobile Synthesiser for the Blind*;[51] *The Online News Reader*[52] (that gathers and reads news in a synthetic voice in Lithuanian). There is also a free speech synthesis application[53] for converting text into speech.

The Lithuanian language needs more annotated speech databases, which calls for concerted efforts to build speech databases for different fields, dialects, age groups, sound environments (among other things), and to make them available to the public (State Commission of the Lithuanian Language, 2020).

Processing, Understanding, and Generating Natural Language

The various projects that have been implemented in Lithuania have produced key open-source tools for the basic analysis of digital texts in the Lithuanian language, such as a segmentor, a lemmatiser, a morphological analyser, a part of speech tagger, a syntactic parser, a spellchecker, a text normaliser, a solution for the advanced search for Lithuanian text indices, and so on (State Commission of the Lithuanian Language, 2020).

On top of that, there are several open-source language identification and semantic analysis solutions: sentiment analysis,[54] the hate-speech recognition tool,[55] the automatic document summary service for Lithuanian documents,[56] the identified entity recognition tool (State Commission of the Lithuanian Language, 2020).

When it comes to generating natural language, it has to be admitted that Lithuania is only making its first steps in this area of language technology.

4.3 Projects, Initiatives and Stakeholders

National Programs for LT/AI

Due to the limited number of speakers, the Lithuanian government and other state institutions support several programs to promote a range of linguistic research and dissemination. The following are valid documents and programs on language technology policies in Lithuania:

- *The Guidelines for the Development of the Lithuanian Language in the Digital Media and Progress in Language Technologies for 2021 – 2027*,[57] issued by the State Commission of the Lithuanian Language. The aim of the guidelines is "to ensure the full functioning and use of the Lithuanian language in the digital environment and the progress of its Lithuanisation, also to promote the development of technologies adapted to the Lithuanian language, and to improve the quality of services to the public that are based on such technologies" (State Commission of the Lithuanian Language, 2020).

[48] https://www.tilde.lt/snekos-technologijos
[49] https://liepa.rastija.lt/Tartuvas/Naujienos?_ga=2.104838787.241622064.1643131617-1654665649.1643131617
[50] https://raštija.lt/liepa/paslaugos-vartotojams/sintezatorius-akliesiems/
[51] https://raštija.lt/liepa-2/paslaugos-vartotojams/mobilusis-sintezatorius-akliesiems/
[52] https://raštija.lt/liepa-2/paslaugos-vartotojams/interneto-naujienu-skaitytuvas/
[53] https://www.tilde.lt/snekos-technologijos
[54] https://ekalba.lt/nuomoniu-analize/
[55] http://hatespeech.vdu.lt
[56] https://www.semantika.lt/Analysis/Summary
[57] https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/911407f20ee911ebbedbd456d2fb030d

- *The Lithuanian Artificial Intelligence Strategy: A Vision for the Future* (2018),[58] issued by the Ministry of the Economy and Innovation of the Republic of Lithuania. The aim of the strategy is "for Lithuania to become a regional leader on the basis of the existing resources, experience and potential. It aims to increase Lithuania's competitiveness among the EU countries and to ensure its successful participation in the global AI ecosystem" (The Ministry of the Economy and Innovation of the Republic of Lithuania, 2018).
- *The Strategy for Lithuania's Advancement* "Lietuva 2030".[59] This strategy outlines the "vision and development priorities of the state as well as the directions of their implementation going forward to 2030. It is the main planning document that must be taken into account in making strategic decisions and drafting state plans or programs" (The Seimas of the Republic of Lithuania, 2012).

National Infrastructures for Language Resources and Technologies

Lithuania has several national technology and language data infrastructures:

- RAŠTIJA.LT[60]: the system of integrated Lithuanian language and written resources, products and services developed by the Institute of Mathematics and Informatics of Vilnius University, providing a knowledge base, search tools, etc.
- CLARIN-LT[61]: the Lithuanian National Consortium (member of CLARIN ERIC) was established in 2015. It currently consists of five research institutions: Vytautas Magnus University (as the coordinator), Kaunas University of Technology, Vilnius University, Mykolas Romeris University and the Baltic Institute of Advanced Technologies.
- E. KALBA[62]: the Lithuanian language resources information system E. KALBA is managed by the Institute of the Lithuanian Language. It features nine monolingual and ten bilingual dictionaries, various files and databases (a dialect archive; a geo-information database of Lithuanian place names, etc.), as well as electronic services (search in the Network of Meanings, E-terms (Ontologies), Sentiment Analysis, etc.).
- SEMANTIKA[63]: the Lithuanian Syntactic and Semantic Analysis Information System (LSSAIS) "is a unique language technology infrastructure and state information system providing speech recognition and text analysis services for the Lithuanian language". Vytautas Magnus University is the manager of the information system.

Consortia, Federations and Projects

- The *European Federation of National Institutions for Language* (EFNIL). Representatives in Lithuania: the Institute of the Lithuanian Language and the State Commission of the Lithuanian Language. The Federation focuses on the languages of the EU Member States and the linguistic diversity of Europe. The goals of this Federation are particularly relevant to the development and advancement of language technology (and AI technology).[64]

[58] https://eimin.lrv.lt/uploads/eimin/documents/files/DI_strategija_LT(1).pdf
[59] https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/TAIS.425517
[60] https://raštija.lt
[61] http://clarin-lt.lt
[62] https://ekalba.lt
[63] https://www.semantika.lt
[64] https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/911407f20ee911ebbedbd456d2fb030d

- The *European Language Resource Coordination* (ELRC) "manages, maintains and coordinates the relevant language resources in all official languages of the EU and CEF associated countries. These activities will help to improve the quality, coverage and performance of automated translation solutions in the context of current and future CEF digital services".[65]
- The *Common Language Resources and Technology Infrastructure* (CLARIN-ERIC) is "a research infrastructure that was initiated from the vision that all digital language resources and tools from all over Europe and beyond are accessible through an online environment for the support of researchers in the humanities and social sciences".[66] CLARIN-LT is a Lithuanian national consortium coordinated by Vytautas Magnus University, which has been a member of CLARIN ERIC.CLARIN-LT since 2014.
- The *European Language Grid* (ELG): "the ELG develops and deploys a scalable cloud platform, providing, in an easy-to-integrate way, access to hundreds of commercial and non-commercial Language Technologies for all European languages, including running tools and services as well as data sets and resources".[67] Since 2019, ELG has been represented in Lithuania by the Institute of the Lithuanian Language.
- *European Language Equality* (ELE): "the primary goal of ELE is to prepare the European Language Equality Programme, in the form of a strategic research, innovation and implementation agenda and a roadmap for achieving full digital language equality in Europe by 2030".[68] Since 2021, ELE has been represented in Lithuania by the Institute of the Lithuanian Language.

Over the period of 2014 – 2020, funding for the implementation of Lithuanian language solutions in the digital space was obtained from the Operational Programme for EU Structural Funds Investments, priority axis 2 (*Promoting the Information Society*), measure titled 'The Lithuanian Language in Information Technologies'. The year 2018 saw the launch of five projects: "Development of Lithuanian Speech-Controlled Services" (LIEPA-2), "Development of the Public Services of the Information System of Syntactic-Semantic Analysis of Lithuanian Texts" (SEMANTIKA 2), "Enhancement and Development of Machine Translation Systems and Localisation Services", "Development of the Information System of Lithuanian Language and Written Resources" (RAŠTIJA 2), and "Development of the Information System of Lithuanian Language Resources" (E. KALBA). A total of 21 public e-services had been developed by the end of 2020. The most active participants of this programme have been research and educational establishments working in association with the industry (State Commission of the Lithuanian Language, 2020).

Apart from the above projects, other language technology projects have been supported by EU structural funds, by the European Commission, or by national funds, and so on.

LT Providers

There are a number of key research and study institutions that contribute to the development of the Lithuanian language in the digital environment. The Institute of the Lithuanian Language conducts research on language data, creates digital resources, various electronic services, and manages the Lithuanian language resource information systems E. KALBA. Vilnius University also develops digital resources, provides various language technology services (e. g. machine translation, etc.), and manages the integrated Lithuanian language and

[65] https://www.lr-coordination.eu
[66] https://www.clarin.eu
[67] https://www.european-language-grid.eu
[68] https://european-language-equality.eu

writing resources, products and services system RAŠTIJA.LT. Vytautas Magnus University creates various language technologies, corpora, and manages the Lithuanian Syntactic and Semantic Analysis Information System SEMANTIKA. Also, there are some key private companies contributing to the development of the Lithuanian language in the digital environment: "Tilde Informacinės Technologijos" provides AI-based development and localisation services for language technologies (speech recognition, synthesis, machine translation, virtual assistants); "Tokenmill" provides development and application of digital text analysis solutions and language technology data sets, specialising in the fields of AI technology, media monitoring, natural language processing and understanding, and language generation; ATEA specialises in the development and implementation of language technology infrastructure solutions; "Algoritmų Sistemos" is an information systems development and implementation company with experience in developing speech technology solutions,[69] etc. (State Commission of the Lithuanian Language, 2020).

5 Cross-Language Comparison

The LT field[70] as a whole has evidenced remarkable progress during the last few 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[71] 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

[69] https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/911407f20ee911ebbedbd456d2fb030d
[70] This section has been provided by the editors.
[71] 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.

- 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 ≥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[72]

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[73] 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

[72] 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.

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

calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[74]

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

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

5.4 Results and Findings

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

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g., German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[75] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

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

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

[74] 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.

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

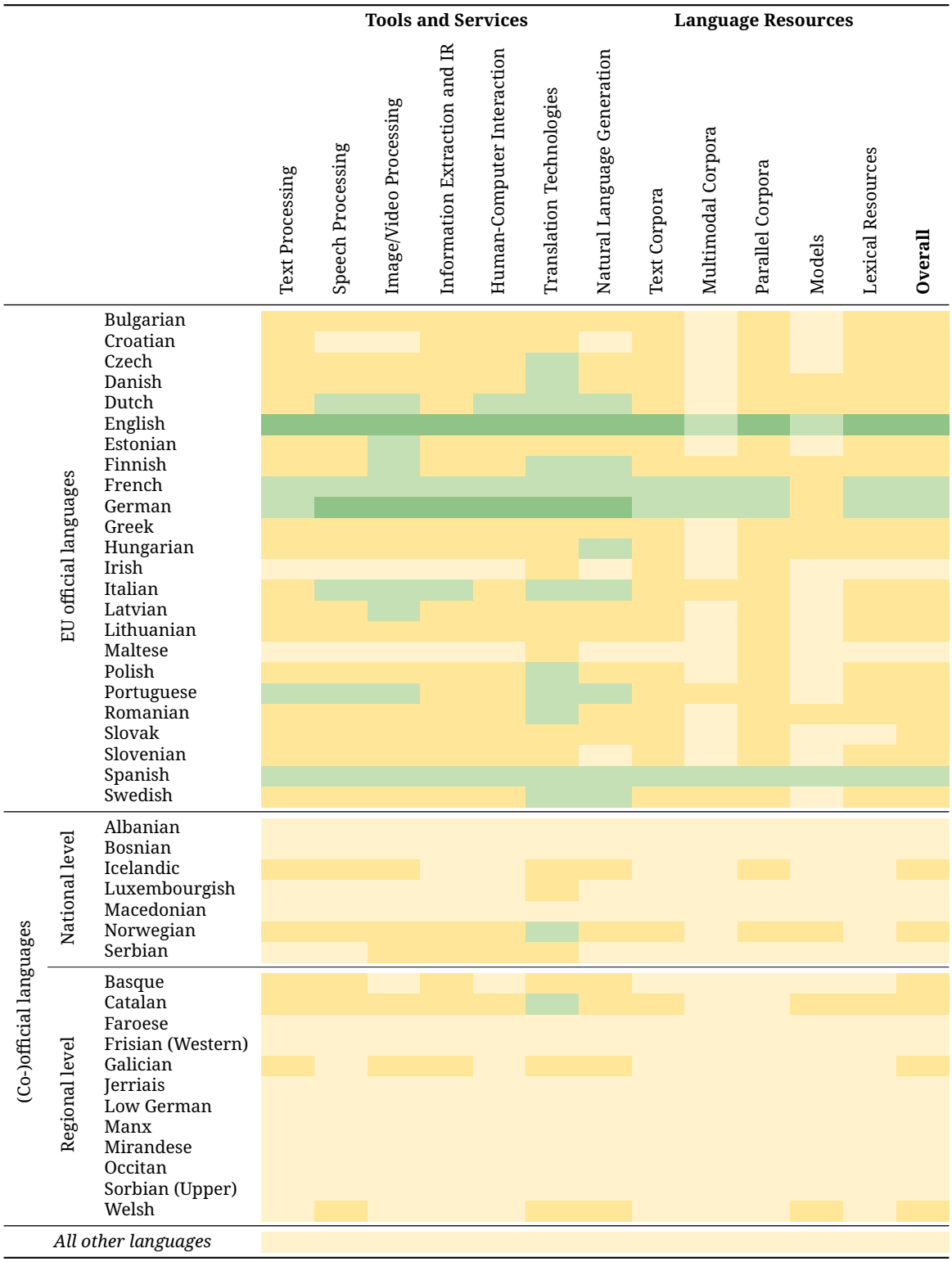

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

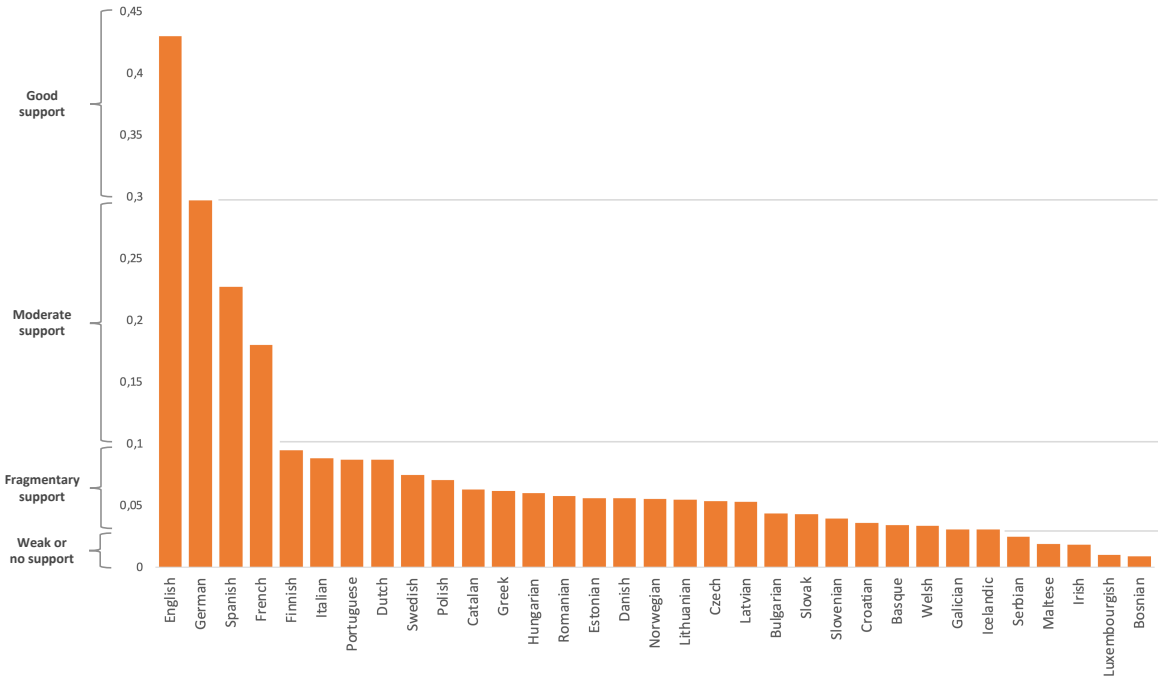

Figure 1: Overall state of technology support for selected European languages (2022)

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

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

6 Summary and Conclusions

Significant progress has been made in adapting the Lithuanian language to the digital environment: a number of digital language resources and basic language analysis tools have been developed, complex online language services have been created, an ontology of the Lithuanian language has been developed and localisation of many computer programs and tools has been achieved. Computer applications relevant to the society have been localised and computer terms standardised. Lithuanian researchers actively participate and cooperate in the mobility activities of international associations. Many Lithuanian language specialists focus on the field of information technology and systematically develop innovative work in this area. Lithuania is very interested in having full access to digital solutions for all

its citizens.

Lithuania still lacks language resources in the electronic environment for faster integration of the Lithuanian language and Information Technologies, as well as standards for the management of such integration. There is a lot of relevant software that has not yet been Lithuanianised and adapted to the needs of society; it is necessary to further ensure the uniform use of computer terms, vocabulary, and phrases in the software, and to continue to take care of the use of the spoken Lithuanian language (speech) in the electronic environment. At a national level, a clear legal framework is essential to ensure equal treatment of research (non-commercial and commercial) innovation based on the automatic extraction and analysis of data from electronic unstructured information sources (texts). Only after a sufficient number of relevant resources reflecting the phenomena of the current Lithuanian language have been gathered, will the development of effective tools for the application of the Lithuanian language in Information Technologies be possible. Particular attention must be paid to adapting to the opportunities and needs of all consumers, so as not to produce social exclusion, which would have implications on society as a whole.

It is important to create Lithuanian interfaces that would directly reduce the social linguistic segregation, promote the legal use of software, and reduce the gap with the EU member states in the use of Information Technology. It is also important to create conditions for the use of the Lithuanian language on computers, computer-controlled devices and computers controlled in the spoken Lithuanian language (speech), to improve the means of computer voice control so that people with disabilities and other vulnerable communities should have unrestricted access to electronic services. Computational linguistics and language technologies, as separate subjects, are not yet established in the Lithuanian tertiary education system. No university offers language technology studies at any level. This needs to change.

In Lithuania, there is a need to increase the competence of specialists working in the field of language technologies and to raise the level of the society's ability to use the opportunities that language technologies have to offer. It is also important to train specialists who know the specifics of language and information technologies, to fund fundamental and applied research, to support scientific and technical infrastructures. Moreover, it is crucial to collect and increase the availability of open, reliable, high-quality, reusable digital language resources and other digital language datasets. There is a need to develop the language technology infrastructure, the application of language technologies in the public sector and public services, teaching and learning institutions, and to develop and improve publicly available IT solutions and tools. Lithuania needs to become even more actively involved in European and other international language technology programs. It is important to update infrastructures with the necessary digital resources, to upgrade and maintain infrastructure hardware, to integrate infrastructures into larger national, European and international language resource platforms, and to ensure the openness of technologies and data stored therein (State Commission of the Lithuanian Language, 2020; Pastor et al., 2017).

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.

Jurgita Jaroslavienė and Albina Auksoriūtė. Innovations and challenges in the digital transformation of the lithuanian language industry. *Language and Economy. Language industries in a Multilingual Europe, EFNIL – European Federation of National Institutions for Language*, 2019.

Jurgita Jaroslavienė and Rita Miliūnaitė. The limitless world of the lithuanian language in the e-language system of digital resources. *The Lithuanian of the World*, 2020.

Rafael Rivera Pastor, Carlota Tarín Quirós, Juan Pablo Villar García, Iclaves, Toni Badia Cardús, Maite Melero Nogués, and Pompeu Fabros University. *Language equality in the digital age. Mother tongue project*. Europe Parlament, 2017.

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.

State Commission of the Lithuanian Language. The guidelines for the development of the lithuanian language language technologies for 2014–2020, 2014.

State Commission of the Lithuanian Language. The guidelines for the development of the lithuanian language in the digital environment and the progress of language technologies for 2021–2027, 2020.

The Ministry of the Economy and Innovation of the Republic of Lithuania. The lithuanian artificial intelligence strategy: A vision for the future. 2018.

The Seimas of the Republic of Lithuania. The strategy for lithuania's advancement "lietuva 2030". 2012.

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.

Daiva Vaišnienė, Jolanta Zabarskaitė, Georg Rehm, and Hans Uszkoreit. *The Lithuanian Language in the digital age*, volume 385. Springer, 2012.

Daiva Vitkutė-Adžgauskienė, Justinas Juozas Dainauskas, Darius Amilevičius, and Andrius Utka. Lietuvių kalbos žodžių tinklas–litwordnet. *Darbai ir dienos. Kaunas, Vilnius: Vytauto Didžiojo universitetas; Versus Aureus, 2015, T. 64*, 2015.