EUROPEAN
LANGUAGE
EQUALITY

D1.22

Report on the Latvian Language

Authors	Inguna Skadiņa, Ilze Auziņa, Baiba Valkovska, Normunds Grūzītis
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	D1.22
Deliverable title	Report on the Latvian Language
Type	Report
Number of pages	29
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	Inguna Skadiņa, Ilze Auziņa, Baiba Valkovska, Normunds Grūzītis
Reviewers	Jaroslava Hlavacova, Maria Giagkou, Andrejs Vasiļjevs
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 © 2022 ELE Consortium

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
AiLab	Artificial Intelligence Laboratory
AI4EU	AI4EU (EU project, 2019-2021)
AMR	Abstract Meaning Representation
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
CEF	Connecting Europe Facility
CLARIN	Common Language Resources and Technology Infrastructure
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
ERIC	European Research Infrastructure Consortium
FAIR	Findability, Accessibility, Interoperability, and Reuse
GPU	Graphics Processing Unit
HCI	Human Computer Interaction (see HMI)
IMCS	Institute of Mathematics and Computer Science
IMCS UL	Institute of Mathematics and Computer Science, University of Latvia
LR	Language Resources/Resources
LT	Language Technology/Technologies
LVK2018	Balanced Corpus of Modern Latvian
ML	Machine Learning
MT	Machine Translation
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
SAFMORIL	CLARIN Knowledge Center for Systems and Frameworks for Morphologically Rich Languages
R&D	Research and Development
TEI	Text Encoding Initiative
TRL	Technology Readiness Level
UD	Universal Dependencies
UL	University of Latvia

Abstract

In this day and age, when our devices are connected in cyberspace, the availability and use of languages and language technology in the digital world varies a lot not only among different languages of the world, but also in Europe, including the official languages of European Union (EU). The European Language Equality (ELE) project provides a survey of language technology support for European languages ten years after the similar exercise was performed and described for 31 European languages in the META-NET White Papers series (Rehm and Uszkoreit, 2012). The reports from the ELE project describe not only the current state of affairs for each of the languages covered, but additionally – and most importantly – they identify the gaps and factors that hinder further development in language technology (LT). Identification of such weaknesses lays the ground for a comprehensive, evidence-based proposal of required measures for achieving Digital Language Equality in Europe by 2030.

This report outlines the state of affairs for the Latvian language, the only official language of Latvia, with about 1.5 million native speakers around the world. Latvian is also spoken as a second language by approximately 500,000 people of other ethnicities.

The necessity of LT support in digital means and the importance of LT for the long-term survival of the Latvian language has always been recognised in policy planning documents. Research and development activities in Latvia are being supported through different EU and national finance instruments and are usually organised around short-term projects. The lack of a dedicated LT programme, however, leads to fragmentation of research and development activities and complicates the development of larger resources and long-term cooperation between institutions.

Since the publication of the META-NET White Paper for Latvian (Skadiņa et al., 2012), notable progress has been made in the development of language resources and tools for Latvian. Today, Latvian is rather well-represented not only by different language resources (digital libraries, text and speech corpora, treebanks, machine-readable lexicons, etc.) but also by core LT, such as spell checkers, morphological analyzers and taggers, named entity recognisers and syntactic parsers, etc. As for more advanced technologies, significant progress has been made in the development of advanced datasets and neural language models, machine translation solutions, speech technologies and technologies for natural language understanding and human-computer interaction. However, when compared to widely spoken and high-resourced languages, there are still significant gaps with respect to solutions that involve deep state of the art natural language understanding and generation, require large and complicated datasets and high performance computing resources.

The Latvian language has a rather stable position in the digital world, and is definitely not in any immediate danger since language resources and LT are in continuous development. However, the situation could change dramatically, if efforts and investments in LT are not increased in the R&D and language policy. Strong national and European support is necessary for further Latvian language research and development activities, including dedicated long-term LT programs, that provide equal support for both research and industrial activities. To narrow the digital divide, there is a pressing urgency for novel techniques, that would bring less resourced languages to a level comparable, to the state of the art results for resource-rich languages. Moreover, close synchronisation between national and international activities is necessary, especially regarding support for research infrastructures and for defining research priorities.

Kopsavilkums

Valodas situācija ir atkarīga ne tikai no tā, cik cilvēku tajā runā, cik grāmatu tajā izdots un filmu uzņemts vai cik televīzijas kanālu tajā pārraida, bet arī no valodas lietojuma informācijas tehnoloģijās un digitālās informācijas telpā. Digitālajā vidē latviešu valodas pozīcijas pēdējā desmitgadē ir būtiski uzlabojušās. Kaut gan kopš META-NET Balto grāmatu publicēšanas 2012. gadā (Skadiņa et al., 2012) latviešu valodas resursi un rīki ir ievērojami uzlabojušies, tomēr daudz plašāk lietotām valodām (piem., angļu, franču vai vācu) tehnoloģiskais nodrošinājums kopumā vēl arvien ir ievērojami attīstītāks.[1]

Šobrīd latviešu valodai izstrādāti ne tikai pamatresursi, bet arī fundamentālas valodas tehnoloģijas. Latvijā galvenā zinātniskā institūcija valodas resursu un tehnoloģiju pētniecībā un izstrādē ir Latvijas Universitātes Matemātikas un informātikas institūts (LU MII), bet vadošais uzņēmums – Tilde. Arvien aktīvāk valodas tehnoloģiju jomā un ar to saistītajos virzienos, piemēram, digitālajās humanitārajās zinātnēs, iesaistās arī citas akadēmiskās institūcijas un komersanti, tiek veidoti jaunuzņēmumi.

Pēdējā desmitgadē ir mērķtiecīgi strādāts pie esošo latviešu valodas korpusu papildināšanas un jaunu korpusu izstrādes. 2012.–2013. gadā tika izveidots pirmais apjomīgais latviešu valodas runas korpuss 100 stundu apjomā (Pinnis et al., 2014). Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa (Levane-Petrova, 2019) apjoms 2022. gadā sasniegs aptuveni 100 milj. vārdlietojumu, savukārt sintaktiski marķētā korpusa (Rituma et al., 2019; Pretkalniņa et al., 2018) apjoms tuvosies 20 tūkst. teikumu. Dažādots brīvi pieejamo specializēto korpusu klāsts: izveidoti vairāki valodas apguvēju korpusi, literārie korpusi, bērnu runas korpuss, emuāru korpuss, subtitru korpuss, Saeimas debašu korpuss, nozarspecifisks medicīnisko diktātu korpuss u.c. Gandrīz visi korpusi ir automātiski morfoloģiski marķēti, bet ir pieejami arī manuāli pārbaudīti, sintaktiski un semantiski marķēti korpusi. Liela daļa no dažādu institūciju veidotajiem brīvpiekļuves korpusiem ir apvienoti Nacionālajā latviešu valodas korpusu kolekcijā un ir pieejami vienotai meklēšanai.[2] Paralēlie korpusi pieejami Opus platformā (Tiedemann, 2016),[3] ELRC-SHARE repozitorijā,[4] Latvijas valsts pārvaldes valodas tehnoloģiju platformā *Hugo.lv* un Tildes datu bibliotēkā.[5]

Īpaša uzmanība ir veltīta latviešu valodas resursu savietojamībai ar citu valodu resursiem. Starp pēdējos gados izveidotajiem latviešu valodas resursiem, kuru marķēšanā ir izmantoti interlingvāli modeļi un kas ir integrēti atvērtās daudzvalodu datu kopās, izceļami šādi: Saeimas debašu korpuss (6,5 milj. vārdu), kurš ir savietojams ar TEI (*Text Encoding Initiative*) un UD (*Universal dependencies*) modeļiem un kurš ir iekļauts CLARIN ERIC daudzvalodu parlamentāro debašu korpusā ParlaMint (Erjavec et al., 2022, 2021); latviešu valodas sintaktiski marķētais korpuss (16 tūkst. teikumu), kurš ir savietojams ar UD gramatikas modeli un ir iekļauts daudzvalodu UD datu kopā (Pretkalniņa et al., 2018); latviešu valodas semantiski marķētais korpuss (Gruzitis et al., 2018a), kurš ir savietojams ar *Berkeley FrameNet* freimu semantikas modeli un tiek gatavots iekļaušanai *Global FrameNet* daudzvalodu datu kopā (vairāk nekā 20 tūkst. freimu).[6]

Vārdnīcu platformā *Tēzaurs.lv*[7] brīvi pieejama lielākā (vairāk nekā 380 tūkst. šķirkļu) mašīnlasāmā latviešu valodas vārdnīca *Tēzaurs* (Spektors et al., 2016), kā arī citas nozīmīgas latviešu valodas vārdnīcas: Mūsdienu latviešu valodas vārdnīca (MLVV; vairāk nekā 45 tūkst. šķirkļu) un Latviešu literarās valodas vārdnīca (LLVV; vairāk nekā 64 tūkst. šķirkļu). Tiek aktīvi izstrādāts Tēzaurs.lv paplašinājums – latviešu valodas leksiskais tīkls *Latvian*

[1] https://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_LV.html
[2] http://korpuss.lv
[3] https://opus.nlpl.eu
[4] https://elrc-share.eu
[5] https://www.tilde.com/products-and-services/data-library
[6] https://globalframenet.org
[7] https://tezaurs.lv

WordNet (Lokmane et al., 2021), kurš ir savietojams un sasaistīts ar *Open Multilingual WordNet* leksisko datubāzi. Dažādas tulkojošās vārdnīcas pieejamas *letonika.lv*.[8] Lielas terminu kolekcijas ir brīvi pieejamas Eiropas Terminoloģijas bankas vietnē[9] un Latvijas Nacionālajā terminoloģijas portālā.[10]

Latviešu valodai ir izstrādāti un attīstīti mūsdienīgi **valodas apstrādes rīki**, kas paredzēti teksta gramatiskajai un semantiskajai marķēšanai, t.sk. vārdšķiru un nosaukto entitāšu noteikšanai, pareizrakstības un gramatikas pārbaudei, runas atpazīšanai un sintēzei, mašīntulkošanai. Dažādi brīvpieejas latviešu valodas apstrādes rīki ir integrēti platformā NLP-PIPE[11] (Znotins and Cirule, 2018), bet pareizrakstības un gramatikas pārbaudītāji un mašīntulki lietotājiem pieejami uzņēmumu Google, Microsoft un Tilde izstrādātajos rīkos.

Aktīva pētniecības joma ir mašīntulkošana (MT). Jaunākie MT risinājumi integrēti Latvijas valsts pārvaldes valodas tehnoloģiju platformā *Hugo.lv*, kas brīvi pieejama ikvienam Latvijas iedzīvotājam un ir īpaši pielāgota valsts pārvaldes dokumentu tulkošanai. Mašīntulkošanas risinājumus piedāvā arī vairāki globālie uzņēmumi. Sabiedrības "Tilde" radītās MT sistēmas uzrādījušas labus rezultātus gan pētnieku rīkotajās sacensībās (Pinnis et al., 2017, 2018; Bojar et al., 2017, 2018), gan komercproduktos ES dalībvalstīm. Šie rezultāti ļāvuši izstrādāt ES Padomes prezidentūras tulkotāju, kas ir izmantots jau 8 valstīs (Pinnis et al., 2020). Tomēr dažādām jomām specifisko apmācību datu trūkums joprojām ierobežo domēnspecifisku MT sistēmu izstrādi mazākām valodām, t.sk. latviešu valodai.

Pietiekami apjomīga (100 stundu) runas korpusa izveide ļāva strauji attīstīt latviešu valodas runas tehnoloģijas, t.sk. izstrādāt plaša lietojuma runas atpazīšanas sistēmas (Znotins et al., 2015; Salimbajevs and Ikauniece, 2017). Publiski pieejams universālais latviešu valodas runas sintezators un runas atpazinējs Tildes Balss,[12] kā arī Google un Microsoft piedāvātie latviešu valodas runas atpazīšanas un sintēzes mākoņpakalpojumi. Tiek veidoti sintezētas balss risinājumi ar emocionālu izteiksmi (Nicmanis and Salimbajevs, 2021). Tiek izstrādātas arī nozarspecifiskas runas atpazīšanas sistēmas, piemēram, LU MII, sadarbojoties ar Rīgas Austrumu klīnisko universitātes slimnīcu izstrādājis medicīnisko diktātu automatizētas transkribēšanas sistēmu (Gruzitis et al., 2022).

Latviešu valodai ir izstrādāti arī dažādi specializēti virtuālie asistenti. Daudzi no tiem tiek izmantoti sabiedrisko pakalpojumu un valsts sektorā, piemēram, Latvijas Bankā, Valsts ieņēmumu dienestā (Skadins et al., 2020).

Dažādu institūciju izstrādātie latviešu valodas resursi un rīki tiek pakāpeniski iekļauti Eiropas pētniecības un valodas tehnoloģiju infrastruktūru CLARIN un ELG repozitorijos.

Kaut gan pēdējā desmitgadē daudz kas ir paveikts, ir arī būtiski trūkumi. Piemēram, trūkst gan vienvalodas, gan daudzvalodu datu. Sarunvalodas korpusi, jautājumu-atbilžu datu kopas, zināšanu bāzes, nozarspecifiskie korpusi ir nelieli vai nav pieejami. Tāpat nav brīvi pieejamu runātās valodas un multimodālo valodas resursu un nozarspecifisku paralēlo un daudzvalodu korpusu specializētu MT sistēmu apmācībai. Ir pārāk maz brīvpieejas tekstu korpusu, kas ļautu apmācīt tādus lielos neironu valodas modeļus kā GPT-3. Īpaši pietrūkst attīstītāku dabiskās valodas sapratnes un sintēzes tehnoloģiju, t.sk. runātās valodas un multimodālo risinājumu.

Lai latviešu valoda arī turpmāk būtu dzīvotspējīga attīstītā digitālajā pasaulē, tai jābūt pieejamiem atbilstošiem resursiem un tehnoloģiskajiem risinājumiem. Tomēr pētniecības un izstrādes darbs ir fragmentārs un galvenokārt tiek organizēts dažādos īstermiņa projektos, kas sarežģī lielāka apjoma resursu izstrādi un uzturēšanu ilgtermiņā. Tikai sistemātisks un mērķtiecīgs valsts atbalsts ilgtspējīgiem pētījumiem valodas resursu un rīku izveidē un uzturēšanā var nodrošināt valodu līdztiesību ikdienas lietojumā un digitālajā vidē.

[8] https://www.letonika.lv/groups/default.aspx?g=2

[9] https://eurotermbank.com

[10] https://termini.gov.lv

[11] http://nlp.ailab.lv

[12] https://www.tilde.lv/tildes-balss

1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. and it seeks to not only delineate the current state of affairs for each of the European languages covered in this series, but to additionally – and most importantly – identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

To this end, more than 40 research partners and 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.

Section 2 introduces the reader to the Latvian language, its status, number of speakers and dialects. It provides an overview of orthography, grammar, punctuation principles and phonetics. Some facts about language use in the digital sphere are also provided. Section 3 introduces readers to the concept of language technology (LT) and its main application areas. The content of this chapter is similar for all language reports of the ELE project, but includes some adaptations from authors of this deliverable. Section 4 provides a high-level overview of Language Technology for Latvian. The section starts with an overview of language resources (corpora, lexical resources and models) and tools (text analysis tools, tools for natural language understanding and generation technologies, machine translation solutions, speech technologies and technologies for human-computer interaction) that are available for the Latvian language. Then, actual information about national programs and policy planning documents, research infrastructures, recent projects and initiatives are summarised. Finally, an overview of language technology providers, researchers and technology developers is presented. Section 5 compares a number of languages investigated by the ELE project with respect to their available resources in the catalogue of the European Language Grid. Finally, Section 6 summarises the findings of this report: strengths and weaknesses, well supported and less supported LT application areas, and main gaps that a large-scale LT R&D programme should try to fill, in order to increase the DLE score for the Latvian language.

The report has been developed by the European Language Equality (ELE) project[13]. 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 Latvian Language in the Digital Age

2.1 General Facts

Latvian is the official language of the Republic of Latvia. This is stipulated by Article 4 of the Constitution of the Republic of Latvia (Satversme) and Article 3 of the State Language Law. When Latvia joined the European Union in 2004, Latvian also became an official language of the European Union. There are approximately 1.5 million native speakers of Latvian, of which, 1.38 million live in Latvia, while the rest live in the United States, Australia, Canada, the United Kingdom, Germany, Lithuania, Estonia, Sweden, Russia, and other countries. Latvian is spoken as a second language by approximately 500,000 people of other ethnicities.[14]

[13] 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.

[14] Latvian population statistics from the Latvian Language Agency Website: https://valoda.lv/valsts-valoda

The Latvian language is the language used in public communication, it is the language taught in schools, the language of public names, signs and writings, the language of the institutions and it is used in work and business environments and services.[15]

Data of the Central Statistical Bureau of Latvia (CSB) show that by the end of 2017 Latvian was the mother tongue of 60.8% of the country's resident population. The statistics for the population mother tongue is similar to that for ethnic composition – at the beginning of 2018 Latvians accounted for 62.2% of the population. In comparison, 36% of Latvia's citizens are native speakers of Russian and 3.2% are native speakers of other languages, e. g. Belorusian, Polish, Ukrainian, Lithuanian.[16]

Latvian has three **dialects**: the Central dialect, Livonic dialect, and High Latvian dialect (Vanags, 2021). The Central dialect is spoken in central Vidzeme (the Vidzeme Central subdialects), Zemgale (the Semigallic subdialects), and in southern Courland or Kurzeme (the Curonic subdialects). The Livonic dialect is spoken in northwestern Vidzeme (the Vidzeme Livonic subdialects) and in northern Courland (the Kurzeme Livonic subdialects or the Tamian subdialects). The Standard Latvian developed primarily based on the Vidzeme Central and Semigallic subdialects (Kalnaca and Lokmane, 2021).

The Latvian language uses the phono-morphological basis of orthography. Latvian orthography almost fully corresponds to the pronunciation. The present-day Latvian **orthography** basis is the Latin script, developed by the Knowledge Commission of the Riga Latvian Association in 1908 and introduced by law from 1920 to 1922 in the Republic of Latvia. Today, the Latvian standard alphabet consists of 33 letters. Some Latvian letters are written with diacritical marks. Macron indicates vowel length. The letters Č, Š and Ž, marked with corona, are pronounced [tʃ], [ʃ] and [ʒ] respectively. The letters Ģ, Ķ, Ļ and Ņ are written with a cedilla or a small comma placed below (or, in the case of the lowercase G, above). They are modified (palatalised) versions of G, K, L and N and represent the sounds [ɟ], [c], [ʎ] and [ɲ]. Latvian orthography also uses digraphs Dz, Dž and IE.

Latvian **punctuation** is based on the grammatical punctuation principle, which means that punctuation marks mainly indicate the grammatical link and division between the text and sentence parts. According to this rule, punctuation marks are used to separate sentences, parts of a compound sentence, equal parts of a sentence, etc. To provide a better representation of nuances in the content of a text or sentence. The grammatical principle is supplemented by the intonational principle (Skadiņa et al., 2012).

From a language **typology** perspective, Latvian has a classic Indo-European (Baltic) system with diverse grammatical inflection and extensive word formation. However, due to regional and historical reasons, Latvian grammar also displays some features more similar to those found in the Finno-Ugric languages (Kalnaca and Lokmane, 2021).

Latvian is a fusional, mainly suffixing language with a rich system of forms and word formation. A distinction is made between inflected (nouns, adjectives, verbs, pronouns, numerals) and non- inflected (adverbs, participles, conjunctions, exclamatives) word classes (Vanags, 2021). Nouns inflect for number and case, adjectives inflect for case, number, gender and definiteness, and verbs inflect for tense, mood, voice and person (Nau, 1998). Word order is relatively free, i. e., pragmatically governed, however, the basic word order is subject verb object (SVO).

There is also a rich system of derivational affixes. The border between inflectional and derivational morphology is not clear-cut (Nau, 1998).

The number of phonemes in standard Latvian is 48: 26 consonant phonemes (/b/, /d/, /f/, /ɡ/, /ɟ/, /x/, /j/, /k/, /c/, /l/, /ʎ/, /m/, /n/, /ɲ/, /ŋ/, /p/, /r/, /s/, /ʃ/, /t/, /v/, /z/, /ʒ/, /ʣ/, /ʤ/, /ʦ/, /ʧ/); 12 wovels – six short vowels (/i/, /e/, /æ/, /ɑ/, /u/, /ɔ/), and six long vowels (/iː/, /eː/, /æː/, /ɑː/, /uː/,

[15] https://valoda.lv/en/state-language/state-language-policy/

[16] https://stat.gov.lv/en/statistics-themes/education/level-education/press-releases/1911-latvian-mother-tongue-608

/ɔː/); 10 diphthongs (/ɑi/, /ui/, /ei/, /ɑu/, /ie/, /uo/, /iu/, /ɔi/, /eu/, /ɔ/), although some diphthongs are mostly limited to proper names and interjections, or only as possitional diphthongs.

Vowel length is phonemic and plays an important role in the language: it distinguishes the lexical and grammatical meaning of the words, for example, *pile* [pile] 'drop' – *pīle* [piːle] 'duck', *māja* [mɑːjɑ] 'house'(nom.sg.) – *mājā* [mɑːjɑː] 'house' (loc.sg.).

Most of the Latvian words are stressed on the first syllable. This holds for the native roots as well as for loanwords, Latvian and foreign proper names. Exceptions to this rule (e. g. ne'viens, pus'otra, all superlative forms of adjective and adverb vis'labākais, vis'tālāk) are rare.

The syllables with the long vowels, diphthongs, and diphthongical combination of vowel and sonorant in the center are subject to certain intonation patterns. In a few areas three patterns of tone or intonation are distinguished: level (also drawling, even) tone, falling tone, and broken tone.

2.2 Latvian in the Digital Sphere

According to the World Bank Data, 89% of the population of Latvia use the Internet. This is an increase of 16% compared to 2012.[17] The number of websites with the country's code (.LV domain names) as top level domain is approximately 136,000.

The language used on the Internet is specific, has certain traditions, and may show characteristics of linguistic impunity (Deksne, 2019).[18]

By compiling statistical data on the behaviour of social network platforms in Latvia, it can be concluded that Latvians are active users of social networks. Facebook is used most frequently by residents of Latvia. According to the latest data of the company NapoleonCat,[19] in August 2021, the number of Facebook users in Latvia reached approximately 1.24 million, which is approximately 67% of the total population of Latvia. The highest participation rate is in the age group between 25 and 34 years of age. Of all users, 56.3% are women, and 43.7% men. The number of Instagram users, on the other hand, is twice as small as Facebook. In Latvia, it is used by about 645 thousand people, of which 58.1% are women and 41.9% are men. The largest number of users is also in the age group from 25 to 34 years (31.6%).

3 What is Language Technology?

Natural language[20] 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 specialised fields known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or, 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 largely multidisciplinary, it combines linguistics, computer science (and notably AI), mathematics and psychology among others.

[17] https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2020locations=LVstart=1990view=chart

[18] This Latvian tweet analysis by (Deksne, 2019) identified several groups of deliberate errors: words without diacritic marks, dropped vowels, new compounds, etc.

[19] http://NapoleonCat.com

[20] 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 and 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[21]. 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.

Language Technology 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 (ASR), i. e. the conversion of speech signal into text, and Speaker Recognition.

- **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,

[21] Main results for this period in Latvia are documented by Spektors (2001), Milčonoka et al. (2004), Vasiļjevs et al. (2004) and Skadiņa (2021)

text re-writing, simplification and generation of questions are some example applications of NLG.

- **Human-Computer Interaction (HCI)** 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). Popular applications within this area are conversational agents (better known as chatbots).

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

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[22]

4 Language Technology for Latvian

Since the publication of the META-NET White Paper for Latvian in 2012 (Skadiņa et al., 2012), the progress and key achievements in the field of language technology have been regularly updated and summarised through the Baltic HLT conferences and some other related events (Vasiljevs and Skadina, 2012; Skadina et al., 2016; Skadina, 2018). These publications demonstrate significant progress in the development of language resources and tools for Latvian, particularly with respect to the creation of advanced datasets and language models, machine translation solutions, speech technologies and technologies for natural language understanding and human-computer interaction.

4.1 Language Data

Text corpora have been developed for the Latvian language already for several decades. Already in 2012, when META-NET White Papers series were published, monolingual written Latvian language corpora were rather well represented, whilst availability of parallel corpora was weak. Moreover, speech corpora for Latvian were not available.

Today most of the open-access monolingual corpora are listed on the *Korpuss.lv* website. Modern Latvian is primarily represented through the Balanced Corpus of Modern Latvian LVK2018 (Dargis et al., 2020b; Levāne-Petrova and Darģis, 2018), which is being extended to 100 million words. For a balanced subset of LVK2018 (FullStack-LV (Gruzitis et al., 2018b;

[22] 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).

Grūzītis et al., 2019), 12.5 thousand sentences), syntactic and semantic annotation layers have been added: named entities, co-references, Universal Dependencies (UD), FrameNet and PropBank annotations, as well as Abstract Meaning Representation. The FullStack-LV multilayer corpus is being enhanced and extended through successive projects, aiming at 20k annotated sentences. Notably, the latest release of the UD layer[23] contains nearly 16k sentences. It should also be noted that the UDLV treebank (Rituma et al., 2021) has been already classified as a big treebank in the CoNLL 2017 and 2018 shared tasks on UD parsing.[24]

Many **parallel corpora** are openly accessible from the Opus platform (Tiedemann, 2016)[25] and the ELRC-SHARE repository[26]. Bilingual and multilingual corpora are also stored at Tilde Data Library (Rozis and Skadiņš, 2017).[27] Tilde Data Library includes 12.35 billion parallel sentences and 23.85 billion monolingual sentences in 124 languages. Part of this content is publicly available from the ELRC and ELG platforms, while some of them are also browsable through *hugo.lv* – the Latvian State Administration Language Technology Platform. However, domain-specific parallel corpora that would allow training and fine-tuning domain-specific MT engines are lacking. For instance, more technical or narrower domains, such as mathematics, physics, chemistry, biology, culture (such as arts, sports, music, etc.), etc. are not well represented in parallel data and can also be scarce in monolingual data, which are used in MT to generate synthetic training data through back-translation (Sennrich et al., 2016).

The first Latvian **speech corpus** was created in 2012–2013 (Pinnis et al., 2014). The corpus contains 100 hours of transcribed speech, which was a good starting point for the development of speech recognition solutions for Latvian. However, access to this speech corpus is limited, and currently the only open-access Latvian speech corpora are LaRKo[28] and Common Voice Latvian[29], each of them containing about 8 hours of annotated speech data. In addition, several domain specific speech corpora (e. g. a medical domain speech corpus (Dargis et al., 2020a)) are currently under development.

Multimodal corpora are still not available for Latvian, although the development of a sign language corpus is planned in the National Research Programme "Letonika – Fostering a Latvian and European Society" project "Research on Modern Latvian Language and Development of Language Technology". In this project, a balanced open-access speech corpus of at least 100 hours will also be created, as well as a quality speech corpus for text-to-speech synthesis.

Latvian digital **lexical resources** are also being developed for a long time. Today, *Tezaurs.lv* is the largest open lexical dataset and on-line dictionary for Latvian (Spektors et al., 2016, 2019). The dictionary is popular not only among researchers, but also widely used by the general public – translators, journalists, students and many others, receiving more than 80,000 requests per day. It is regularly updated, and currently contains more than 380,000 lexical entries that are compiled from more than 300 sources. Another important lexical resource, the Latvian WordNet, is currently under development (Lokmane et al., 2021) and is being integrated with *Tezaurs.lv*.

Different lexicons (mostly bilingual) are available from the *letonika.lv* portal.[30] It contains electronic dictionaries for widely used language pairs (Latvian and English, French, German and Russian), as well as dictionaries of the languages of the Baltic countries: Latvian and Lithuanian, Latvian and Estonian.

[23] https://github.com/UniversalDependencies/UD_Latvian-LVTB
[24] http://universaldependencies.org/conll18/
[25] https://opus.nlpl.eu
[26] https://elrc-share.eu
[27] https://www.tilde.com/products-and-services/data-library
[28] http://korpuss.lv/id/LaRKo (Auziņa et al., 2014)
[29] https://commonvoice.mozilla.org/en/datasets
[30] https://www.letonika.lv/groups/default.aspx?g=2

Two large **terminology collections** are freely available for browsing through the European Terminology Bank (Eurotermbank) website[31] and the Latvian national terminology portal.[32] Today (December, 2021), Eurotermbank (Rirdance and Vasiljevs, 2006) contains about 3.5 million entries (14.5 million terms) from 463 collections in 44 languages.

Finally, several BERT-based language models for Latvian are created and used for named entity recognition, parsing and intent detection (Znotins and Barzdins, 2020; Vīksna and Skadiņa, 2020).

4.2 Language Technologies and Tools

Various basic **text analysis** tools, such as tokenisers, morphological analyzers and taggers, spelling checkers, syntactic parsers, named entity recognisers, etc., are available for Latvian. Spelling and grammar checking tools are available for users through Microsoft and Tilde products, while various open-source Latvian NLP tools are integrated into NLP-PIPE[33] – a modular pipeline for text tokenisation and sentence splitting, morphological tagging, named entity recognition, syntactic parsing, semantic parsing, etc. (Znotins and Cirule, 2018; Znotiņš, 2015). In addition to the above mentioned text analytic tools, several sentiment analysis tools have been created as well.

Regarding **natural language understanding (NLU) and generation (NLG)**, experiments with the interlingual FrameNet, Abstract Meaning Representation (AMR) and Grammatical Framework models for Latvian, English and Swedish (Gruzitis et al., 2017; Ranta et al., 2020) demonstrate the potential of combining machine learning and knowledge-based approaches for state-of-the-art semantic parsing and semantically precise and controllable language generation for both highly-resourced and less-resourced languages.

With respect to **machine translation** (MT), the situation has changed a lot when compared to 2012. Today several machine translation solutions are available from global companies.[34] Language technology company Tilde[35] provides customised MT solutions for complex, highly inflected languages, particularly smaller European languages.[36] MT systems developed by Tilde have been recognised among the best systems for four consecutive years (2017-2020) at WMT international news translation shared tasks (Pinnis et al., 2017, 2018; Bojar et al., 2017, 2018). These results allowed Tilde together with partners to develop EU Council Presidency Translator, which has been used already in 8 countries (Pinnis et al., 2020). However, lack of domain-specific training data still limits development of domain-specific MT engines for smaller languages like Latvian.

For many years Latvian was not so well represented in **speech technologies** due to the lack of speech recognition tools. Shortly after the transcribed 100-hour corpus of spoken Latvian was created, several speech recognition systems were developed (Znotins et al., 2015; Salimbajevs and Ikauniece, 2017). Today, the output of these systems is comparable to the state of the art. A general-purpose Latvian speech synthesiser and speech recogniser by Tilde are publicly available.[37]

Several task-oriented **virtual assistants** can communicate in Latvian, helping users to find answers for particular questions. Virtual assistants are also used by public services. For example, *Hugo.lv* (Skadins et al., 2020) lists more than 10 virtual assistants for different public services, including the Bank of Latvia, the State Revenue Service, The Latvian State Radio and

[31] https://eurotermbank.com

[32] https://termini.gov.lv

[33] http://nlp.ailab.lv

[34] e. g. Google Translate (https://translate.google.lv), Microsoft Translator (https://translator.microsoft.com), Yandex Translate (https://translate.yandex.com)

[35] https://tilde.com/

[36] https://translate.tilde.com

[37] https://www.tilde.lv/tildes-balss

Television Centre and many others. Technologies allowing users to create a virtual conversational agent that can understand textual or voice inputs, identify the user's intent (Kapočiūtė-Dzikienė et al., 2021; Balodis and Deksne, 2019) and deliver response via text, visual media, or voice are provided through *tilde.ai* conversational AI platform.

4.3 Projects, Initiatives, Stakeholders

Research and development activities in Latvia are mostly supported through several instruments: State Research Programmes, EU Structural Funds Programmes (in particular, Competence Centre projects and Practical oriented research projects), grants of the Latvian Science Council, EU Horizon 2020, Horizon Europe and CEF Programmes[38].

National programmes

The necessity of language technology support in digital means and importance of language technologies for the long-term survival of the Latvian language has been always recognised at the policy planning documents. However, there is no dedicated language technology program in Latvia. As a result, research and development activities in human language technologies, as well as creation and long-term maintenance activities related to language resources and tools, are fragmented and not always sufficiently supported.

Several recent policy planning documents stress the importance of support for the Latvian language in digital means:

- The State Language Policy Guidelines for 2021-2027[39] includes activities related to the creation and further development of Latvian language resources and tools. The guidelines are implemented through the three year State Research Programme "Letonika – Fostering a Latvian and European Society"
- Digital Transformation Guidelines for 2021-2027[40] include targeted action line "Machine Translation and Language Technologies" with a vision to enable Latvian citizens to access European Digital Space in their native language and to support the Latvian language with the most important language resources for sustainable development and wide use in digital services.
- Information report "On the development of Artificial Intelligence solutions"[41] lists several future directions of action related to the development of AI-based language technologies – machine translation, speech technologies, inclusive technologies and terminology databases.
- Latvia's Recovery and sustainability plan[42] includes activities related to language technologies, in particular, the plan includes activities related to high level skills in language technologies.

[38] According to the data of the Ministry of Education and Research (https://www.clarin.lv/images/IZMprez_OpenScience-CLARIN.pdf) in 2014-2018 total funding for LT R&D activities was 4.75 million euros: 30% Competence Centre projects, 29% Horizon 2020 projects, 29% Practical oriented research projects, 10% fundamental and applied research program

[39] https://likumi.lv/ta/id/325679-par-valsts-valodas-politikas-pamatnostadnem-2021-2027-gadam

[40] https://likumi.lv/ta/id/324715-par-digitalas-transformacijas-pamatnostadnem-20212027-gadam

[41] http://tap.mk.gov.lv/lv/mk/tap/?pid=40475479

[42] https://likumi.lv/ta/id/322858-par-latvijas-atveselosanas-un-noturibas-mehanisma-planu

Research Infrastructures

Since June, 2016 Latvia is member of CLARIN ERIC (Skadiņa et al., 2020). The coordinating center of CLARIN Latvia[43] is the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL). CLARIN Latvia focuses on Latvian (and Latgalian) language resources, but not excluding other languages, in particular morphologically rich languages. CLARIN Latvia is supported with rather small funding for consortium building and infrastructure building activities. Interruption in funding for six years and lack of sufficient funding currently are the main reasons why the CLARIN-LV repository[44] was set up only in March, 2020.

CLARIN Latvia actively participates in CLARIN ERIC activites, such as CLARIN Resource Families[45], Teaching with CLARIN[46] and ParlaMint (Erjavec et al., 2021). CLARIN Latvia is also a member of CLARIN Knowledge Center for Systems and Frameworks for Morphologically Rich Languages SAFMORIL[47], which brings together researchers and developers in the area of computational morphology and its application in language processing.[48]

Recent projects

During last five years development of Latvian language technologies has been supported through different Horizon 2020 and CEF projects on machine translation[49], human-centred AI [50], speech technologies [51] and activities for support digital language equality[52].

Five large projects have been supported through the Industry-Driven Research Programme of the European Regional Development Fund. These projects focus on the development of basic language resources and tools for deep learning and natural language understanding[53], on the innovative application of speech technologies for multilingual meeting management[54] and the transcription of medical records[55], as well as the development of cognitive intelligent virtual assistants.[56]

Since 2011 more than 15 language technology projects have been implemented with support from the IT Competence Centre Programme[57]. The Competence Centre supported the creation of the first transcribed Latvian speech corpus, followed by several speech projects. More than five projects address machine translation problems, while several projects are related to intelligent virtual assistants and human-computer interaction.

[43] https://www.clarin.lv/en-us/

[44] https://repository.clarin.lv

[45] https://www.clarin.eu/resource-families

[46] https://www.clarin.eu/content/teaching-clarin

[47] https://www.kielipankki.fi/safmoril/

[48] Other members of SAFMORIL are University of Helsinki and CSC (FIN-CLARIN), University of Tromsø (CLARINO), and Vytautas Magnus University (CLARIN-LT)

[49] E.g., MT4All (http://ixa2.si.ehu.eus/mt4all/project), NLTP, NTEU (Bié et al., 2020), CEF Presidency Translator projects, FedTerm

[50] E.g., StairwAI (https://stairwai.nws.cs.unibo.it/), Intelcomp (https://intelcomp.eu/), HumanE-AI-Net (https://www.humane-ai.eu/), AI4EU (https://www.ai4europe.eu/)

[51] E.g., COMPRISE (Skadiņš and Salimbajevs, 2020), SUMMA (http://summa-project.eu/), SELMA (https://selma-project.eu/

[52] E.g., ELG (Rehm et al., 2020b), ELE (Rehm et al., 2020a), MAPA (Ajausks et al., 2020)

[53] Project ”Neural Network Modelling for Inflected Natural Languages” (1.1.1.1/16/A/215) and project “Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian” (1.1.1.1/16/A/219)

[54] Project “AI Assistant for Multilingual Meeting Management” (1.1.1.1/19/A/082)

[55] Project “Latvian Speech Recognition and Synthesis for Medical Applications” (1.1.1.1/18/A/153)

[56] Project “Multilingual Artificial Intelligence Based Human Computer Interaction” (1.1.1.1/18/A/148)

[57] https://www.itkc.lv

LT providers

Systematic research and development activities in the field of language technologies are mostly carried out by two institutions in Latvia: The Institute of Mathematics and Computer Science at the University of Latvia (IMCS UL), and the company Tilde.[58]

The Artificial Intelligence Laboratory[59] (AiLab) at IMCS UL is the leading language technology research group in Latvia, focusing on natural language understanding (NLU) and generation (NLG) (Auzina et al., 2021). Although AiLab primarily focuses on Latvian, it has successfully participated in international NLU and NLG evaluation campaigns on well-resourced languages as well. AiLab also actively participates in the Universal Dependencies (UD), Multilingual FrameNet, Global WordNet and other international initiatives through the development of advanced language resources: Latvian UD Treebank, Latvian FrameNet, Latvian WordNet, etc. In 2016, the laboratory researchers achieved the best result in the shared task of English-to-AMR parsing in the prestigious "SemEval" competition, while in the 2017 competition – the best result in the shared task on AMR-to-English generation.

Tilde[60] is a leading European language technology company, specialising in custom machine translation systems, intelligent virtual assistants, speech technologies and online terminology services. Tilde has experience in developing high-demand cloud-based and desktop language technologies for complex, highly inflected languages, particularly smaller European languages. The technologies created by the company are also used outside Latvia, for example, Tilde's neural machine translation systems have been supporting translation efforts for the EU Council presidencies in Estonia, Bulgaria, Austria, Romania, Finland, Croatia and Germany. The Latvian Academy of Sciences recognised the neural machine translation solution developed by Tilde as one of the most significant achievements of Latvian science in 2018.

Latvian language technologies are also being developed by global companies: *Google* provides machine translation, speech synthesis and speech recognition services; *Microsoft* provides proofing tools, machine translation and speech synthesis services; *Facebook* provides state-of-the-art models for multilingual ASR (including Latvian), and also supports translation into Latvian in its applications.

Latvian language resources are being developed and related research activities are also performed at other research institutions in Latvia: The Institute of Latvian Language at University of Latvia, The Livonian Institute at University of Latvia, The Institute of Literature, Folklore and Art at The University of Latvia, The National Library of Latvia and the Latvian Language Agency, Liepaja University, Ventspils University College and Rezekne Academy of Technology.

5 Cross-Language Comparison

The LT field[61] 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).

[58] https://enciklopedija.lv/skirklis/106524-datorlingvistika-Latvijā

[59] http://ailab.lv/en/

[60] https://tilde.com/

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

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[62] broadly categorised into a number of core LT application areas:
  - Text processing (e. g., part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g., search and information mining)
  - Translation technologies (e. g., machine translation, computer-aided translation)
  - Natural language generation (e. g., text summarisation, simplification)
  - Speech processing (e. g., speech synthesis, speech recognition)
  - Image/video processing (e. g., facial expression recognition)
  - Human-computer interaction (e. g., tools for conversational systems)
- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

5.2 Levels of Technology Support

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

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in $\geq$3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in $\geq$10% and <30% of the ELG resources of the same type

[62] Tools tagged as “language independent” without mentioning any specific language are not taken into account. Such tools can certainly be applied to a number of languages, either as readily applicable or following fine-tuning, adaptation, training on language-specific data etc., yet their exact language coverage or readiness is difficult to ascertain.

4. **Good support**: the language is present in $\geq$30% of the ELG resources of the same type[63]

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[64] 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.[65]

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

[63] 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.

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

[65] 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.

			Tools and Services							Language Resources					
			Text Processing	Speech Processing	Image/Video Processing	Information Extraction and IR	Human-Computer Interaction	Translation Technologies	Natural Language Generation	Text Corpora	Multimodal Corpora	Parallel Corpora	Models	Lexical Resources	**Overall**
EU official languages		Bulgarian	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨
		Croatian	🟨	⬜	⬜	🟨	🟨	🟨	⬜	🟨	⬜	🟨	⬜	🟨	🟨
		Czech	🟨	🟨	🟨	🟨	🟨	🟢	🟨	🟨	⬜	🟨	⬜	🟨	🟨
		Danish	🟨	🟨	🟨	🟨	🟨	🟢	🟨	🟨	⬜	🟨	🟨	🟨	🟨
		Dutch	🟨	🟢	🟢	🟨	🟢	🟢	🟢	🟨	⬜	🟨	🟨	🟨	🟨
		English	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟢	🟩	🟢	🟩	🟩
		Estonian	🟨	🟨	🟢	🟨	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨
		Finnish	🟨	🟨	🟢	🟨	🟨	🟢	🟢	🟨	🟨	🟨	🟨	🟨	🟨
		French	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟨	🟢	🟢
		German	🟢	🟩	🟩	🟩	🟩	🟩	🟩	🟢	🟢	🟢	🟨	🟢	🟢
		Greek	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	⬜	🟨	🟨	🟨	🟨
		Hungarian	🟨	🟨	🟨	🟨	🟨	🟨	🟢	🟨	⬜	🟨	🟨	🟨	🟨
		Irish	⬜	⬜	⬜	⬜	⬜	🟨	⬜	🟨	⬜	🟨	⬜	⬜	⬜
		Italian	🟨	🟢	🟢	🟢	🟨	🟢	🟢	🟨	🟨	🟨	⬜	🟨	🟨
		Latvian	🟨	🟨	🟢	🟨	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨
		Lithuanian	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨
		Maltese	⬜	⬜	⬜	⬜	⬜	🟨	⬜	⬜	⬜	🟨	⬜	⬜	⬜
		Polish	🟨	🟨	🟨	🟨	🟨	🟢	🟨	🟨	⬜	🟨	⬜	🟨	🟨
		Portuguese	🟢	🟢	🟢	🟨	🟨	🟢	🟢	🟨	🟨	🟨	⬜	🟨	🟨
		Romanian	🟨	🟨	🟨	🟨	🟨	🟢	🟨	🟨	⬜	🟨	🟨	🟨	🟨
		Slovak	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	⬜	🟨	⬜	⬜	🟨
		Slovenian	🟨	🟨	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	⬜	🟨	🟨
		Spanish	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢	🟢
		Swedish	🟨	🟨	🟨	🟨	🟨	🟢	🟢	🟨	🟨	🟨	⬜	🟨	🟨
(Co-)official languages	National level	Albanian	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Bosnian	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Icelandic	🟨	🟨	🟨	⬜	⬜	🟨	🟨	⬜	⬜	🟨	⬜	⬜	🟨
		Luxembourgish	⬜	⬜	⬜	⬜	⬜	🟨	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Macedonian	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Norwegian	🟨	🟨	🟨	🟨	🟨	🟢	🟨	🟨	⬜	🟨	🟨	⬜	🟨
		Serbian	⬜	⬜	🟨	🟨	🟨	🟨	⬜	⬜	⬜	⬜	⬜	⬜	⬜
	Regional level	Basque	🟨	🟨	⬜	🟨	⬜	🟨	🟨	⬜	⬜	⬜	⬜	⬜	🟨
		Catalan	🟨	🟨	🟨	🟨	🟨	🟢	🟨	🟨	⬜	⬜	🟨	🟨	🟨
		Faroese	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Frisian (Western)	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Galician	🟨	⬜	🟨	🟨	⬜	🟨	🟨	⬜	⬜	⬜	⬜	⬜	🟨
		Jerriais	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Low German	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Manx	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Mirandese	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Occitan	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Sorbian (Upper)	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
		Welsh	⬜	🟨	⬜	⬜	⬜	🟨	🟨	⬜	⬜	⬜	🟨	⬜	🟨
All other languages			⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜

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

national or regional level in at least one European country and other minority and lesser spoken languages,[66] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

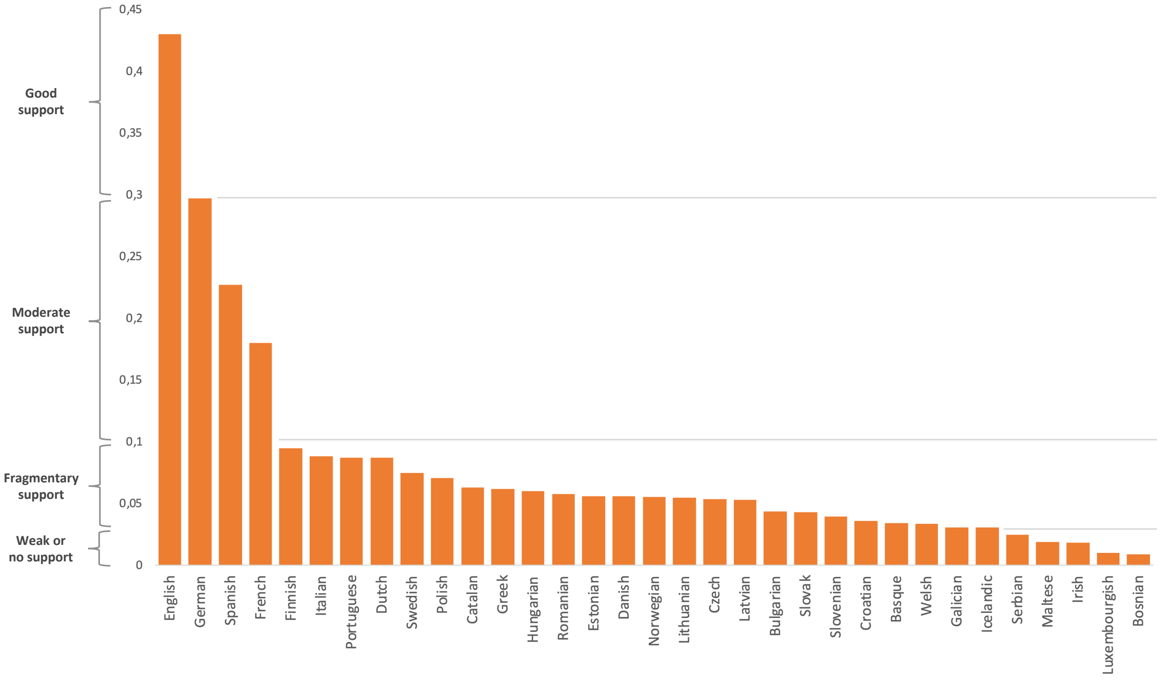

Figure 1: Overall state of technology support for selected European languages (2022)

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

The results of the present comparative evaluation reflect, in terms of distribution and imbalance, the results of the META-NET White Paper Series (Rehm and Uszkoreit, 2012). The complexities of the analyses clearly differ across 2012 and 2022 studies, and as such, a direct comparison between the two studies can therefore not be made. However, we can in-

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

stead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e.g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

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

6 Summary and Conclusions

This report provided a short summary of the current state of the Latvian language in the digital environment – general facts, availability of language data and tools, major projects, initiatives and stakeholders. Since the publication of the META-NET White Papers in 2012, **significant progress** has been made in the development of various language resources and tools for Latvian. Although the Latvian language is represented by a rather small number of speakers and it is often categorised as less-resourced, it is **represented rather well** not only by different **language resources** (digital libraries, text and speech corpora, lexicons, etc.) but also by **core language technologies**, such as spelling checkers, morphological analyzers and taggers, named entity recognisers, syntactic parsers, etc.

Concerning more advanced technologies, Latvian has a **good support for machine translation, speech recognition and synthesis**, while solutions that involve deep state of the art natural language understanding are not so developed.

There are still significant gaps with respect to availability, size and technology readiness level (TRL) of language resources, models and tools, and human, computational and financial resources.

With respect to **language resources**, significant gaps are identified for both monolingual and multilingual data of all forms: written, spoken and multimodal. For example, datasets that represent conversational data, question answering, knowledge bases, informal language or specific domain are small or even unavailable. There are almost no spoken and multimodal open-data or open-access language resources available. Also, domain-specific parallel and multilingual data that would allow training and fine-tuning domain-specific MT engines are insufficient, while the current open-access monolingual text corpora are too small for training massive language models like GPT-3.

Consequently, there is lack of **large pre-trained language models** (both general and domain specific) and lack of benchmarks for specific NLP tasks, e. g. Latvian GLUE or Latvian SQUAD. Creation of such models is limited not only by availability of necessary data but also by **insufficient hardware infrastructure**, which could be solved through significant long-term support for research infrastructures.

The data sharing culture is still developing, partly due to the late implementation of CLARIN, however, recently it has become more acknowledged among DH researchers, for example, through the State research program project “Digital Resources for Humanities: Integration and Development”.[67]

[67] http://www.digitalhumanities.lv/projects/DHVPP-en/

Another important aspect is **IPR and GDPR regulations** that need to be more flexible, allowing wider use of IPR protected data for the development of language technologies and resources in a way that does not harm the interests of the authors.

Overall, similarly to many other languages of Europe, there is insufficient amounts of quality corpora, including monolingual corpora, currently available for Latvian, as well as insufficient computational resources, for training large-scale SOTA language models like the GPT-3 model for English. However, there are resources and competence available for pre-training relatively smaller language models like BERT and GPT-2 and for fine-tuning large pre-trained multilingual models like mT5 and XLS-R for various downstream tasks.

Availability of necessary **human resources** are limited by three factors: the well known deficit of IT specialists and scholars in general, the lack of specialised study programs (modules) and the size of the Latvian population in general. This leads to **gaps and limitations in language technology development**. Although the Latvian LT industry and research groups have demonstrated excellent results in LT adaption for morphologically rich languages (which is not a trivial task), they are less present among leaders in the development of world-class novel language technology solutions.

Finally, **gaps and fragmentation** in research and development activities related to language resources and tools is a result of short, project-based (mostly 2-3 years, sometimes even 1 year, rarely 5 years) research and development funding and disproportion between funding for research (TRL 1-4) and industrial activities (TRL 5-9).[68] With respect to policies/instruments, strong national and international support is necessary for further Latvian language research and development activities, including **dedicated long-term LT programs that provide equal support for both research and industrial activities**. Moreover, close **synchronisation between national and international activities** is necessary, especially, with respect to research infrastructures and research priorities. An instrument for efficient and homogeneous implementation of policies towards Digital Language Equality would be equal international support to national LT research and development communities.

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.

Ēriks Ajausks, Victoria Arranz, Laurent Bié, Aleix Cerdà-i Cucó, Khalid Choukri, Montse Cuadros, Hans Degroote, Amando Estela, Thierry Etchegoyhen, Mercedes García-Martínez, Aitor García-Pablos, Manuel Herranz, Alejandro Kohan, Maite Melero, Mike Rosner, Roberts Rozis, Patrick Paroubek, Artūrs Vasiļevskis, and Pierre Zweigenbaum. The multilingual anonymisation toolkit for public administrations (MAPA) project. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 471–472, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.57.

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.

[68] According to the data of the Ministry of Education and Research (https://www.clarin.lv/images/IZMprez_OpenScience-CLARIN.pdf) in 2014-2018 funding for LT R&D activities was 4.75 million euros, with about 70% (3.5 million) for industrial research

Ilze Auziņa, Roberts Darģis, Kristīne Levāne-Petrova, Kristīne Pokratniece, and Daira Vēvere. Latvian Speech Corpus (LaRKo), 2014. URL http://hdl.handle.net/20.500.12574/22. CLARIN-LV digital library at IMCS, University of Latvia.

Ilze Auzina, Normunds Gruzitis, and Guntis Barzdins. Research and Innovation in Language Technology at the Artificial Intelligence Laboratory. In *The Latvian Academy of Sciences Yearbook*, pages 81–83. Zinātne, 2021. URL https://www.lza.lv/images/Annual_reports/YearBook_2021.pdf.

Kaspars Balodis and Daiga Deksne. Fasttext-based intent detection for inflected languages. *Information*, 10(5), 2019. ISSN 2078-2489. doi: 10.3390/info10050161. URL https://www.mdpi.com/2078-2489/10/5/161.

Laurent Bié, Aleix Cerdà-i Cucó, Hans Degroote, Amando Estela, Mercedes García-Martínez, Manuel Herranz, Alejandro Kohan, Maite Melero, Tony O'Dowd, Sinéad O'Gorman, Mārcis Pinnis, Roberts Rozis, Riccardo Superbo, and Artūrs Vasiļevskis. Neural translation for the European Union (NTEU) project. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 477–478, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.60.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4717. URL https://aclanthology.org/W17-4717.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6401. URL https://aclanthology.org/W18-6401.

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

Roberts Dargis, Normunds Gruzitis, Ilze Auzina, and Kaspars Stepanovs. Creation of Language Resources for the Development of a Medical Speech Recognition System for Latvian. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 135–141. IOS Press, 2020a. doi: 10.3233/FAIA200615. URL https://ebooks.iospress.nl/volumearticle/55536.

Roberts Dargis, Kristine Levane-Petrova, and Ilmars Poikans. Lessons learned from creating a balanced corpus from online data. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 127–134. IOS Press, 2020b. doi: 10.3233/FAIA200614. URL https://ebooks.iospress.nl/volumearticle/55535.

Daiga Deksne. Chat language normalisation using machine learning methods. In *NLPinAI 2019 - Special Session on Natural Language Processing in Artificial Intelligence*, 2019.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Vladislava Grigorova, Michał Rudolf, Andrej Pančur, Matyáš Kopp, Starkaður Barkarson, Steinh͡ór Steingrímsson, Henk van der Pol, Griet Depoorter, Jesse de Does, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, María Calzada Pérez, Luciana D. de Macedo, Ruben van Heusden, Maarten Marx, Çağrı Çöltekin, Matthew Coole, Tommaso Agnoloni, Francesca Frontini, Simonetta Montemagni, Valeria Quochi, Giulia Venturi, Manuela Ruisi, Carlo Marchetti, Roberto Battistoni, Miklós Sebők, Orsolya Ring, Roberts Darģis, Andrius Utka, Mindaugas Petkevičius, Monika Briedienė, Tomas Krilavičius, Vaidas Morkevičius, Sascha Diwersy, Giancarlo Luxardo, and Paul Rayson. Multilingual comparable corpora of parliamentary debates ParlaMint 2.1, 2021. URL http://hdl.handle.net/11356/1432. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 2022. doi: 10.1007/s10579-021-09574-0.

Normunds Gruzitis, Didzis Gosko, and Guntis Barzdins. Rigotrio at semeval-2017 task 9: Combining machine learning and grammar engineering for amr parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, 2017. URL http://www.aclweb.org/anthology/S17-2159.

Normunds Gruzitis, Gunta Nespore-Berzkalne, and Baiba Saulite. Creation of Latvian Framenet based on Universal Dependencies. In *Proceedings of the International FrameNet Workshop (IFNW)*, pages 23–27, 2018a. URL http://lrec-conf.org/workshops/lrec2018/W5/pdf/9_W5.pdf.

Normunds Gruzitis, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins, and Peteris Paikens. Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 4506–4513, 2018b. URL http://www.lrec-conf.org/proceedings/lrec2018/pdf/935.pdf.

Normunds Grūzītis, Lauma Pretkalniņa, Baiba Saulīte, Laura Rituma, Gunta Nešpore-Bērzkalne, Pēteris Paikens, Ilze Auziņa, Artūrs Znotiņš, Kristīne Levāne-Petrova, and Roberts Darģis. Full Stack of Latvian Language Resources for NLU, 2019. URL http://hdl.handle.net/20.500.12574/5. CLARIN-LV digital library at IMCS, University of Latvia.

Normunds Gruzitis, Roberts Dargis, Viesturs Lasmanis, Ginta Garkaje, and Didzis Gosko. Adapting Automatic Speech Recognition to the Radiology Domain for a Less-Resourced Language: The Case of Latvian. In *Intelligent Sustainable Systems*, Lecture Notes in Networks and Systems. Springer, 2022.

Andra Kalnaca and Ilze Lokmane. *Latvian Grammar*. University of Latvia Press, 2021.

Jurgita Kapočiūtė-Dzikienė, Askars Salimbajevs, and Raivis Skadiņš. Monolingual and cross-lingual intent detection without training data in target languages. *Electronics*, 10(12), 2021. ISSN 2079-9292. doi: 10.3390/electronics10121412. URL https://www.mdpi.com/2079-9292/10/12/1412.

Kristine Levane-Petrova. Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. *Language: Meaning and Form*, 10:131–146, 2019. doi: 10.22364/vnf.10.12. URL https://www.apgads.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/Valoda-nozime-forma/VNF-10/vnf_10-12_Levane_Petrova.pdf. The Balanced Corpus of Modern Latvian, its role in grammar studies.

Kristīne Levāne-Petrova and Roberts Darģis. Balanced Corpus of Modern Latvian (LVK2018), 2018.

Ilze Lokmane, Laura Rituma, Madara Stade, and Agute Klints. The Latvian WordNet and Word Sense Disambiguation: Challenges and Findings. In *Proceedings of the 7th Biennial Conference on Electronic Lexicography (eLex)*, pages 232–246, 2021. URL https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_13_pp232-246.pdf.

Everita Milčonoka, Normunds Grūzītis, and Andrejs Spektors. Natural language processing at the Institute of Mathematics and Computer Science: 10 years later. In *Proceedings of the first Baltic conference" Human Language Technologies-the Baltic Perspective*, pages 6–11, 2004.

Nicole Nau. *Latvian*, volume 217. Lincom Europa, 1998.

Dāvis Nicmanis and Askars Salimbajevs. Expressive Latvian Speech Synthesis for Dialog Systems. 2021.

Marcis Pinnis, Ilze Auzina, and Karlis Goba. Designing the Latvian speech recognition corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/284_Paper.pdf.

Mārcis Pinnis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics. Tilde's machine translation systems for WMT 2017. In *Proceedings of the Second Conference on Machine Translation*, pages 374–381, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4737. URL https://aclanthology.org/W17-4737.

Marcis Pinnis, Matiss Rikters, and Rihards Krislauks. Tilde's machine translation systems for WMT 2018. In Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 473–481. Association for Computational Linguistics, 2018. doi: 10.18653/v1/w18-6423. URL https://doi.org/10.18653/v1/w18-6423.

Mārcis Pinnis, Toms Bergmanis, Kristīne Metuzāle, Valters Šics, Artūrs Vasiļevskis, and Andrejs Vasiļjevs. A Tale of Eight Countries or the EU Council Presidency Translator in Retrospect. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 525–546, Virtual, October 2020. Association for Machine Translation in the Americas. URL https://aclanthology.org/2020.amta-user.25.

Lauma Pretkalniņa, Laura Rituma, and Baiba Saulīte. Deriving Enhanced Universal Dependencies from a Hybrid Dependency-Constituency Treebank. In *Text, Speech, and Dialogue*, volume 11107, pages 95–105. Springer, 2018. doi: 10.1007/978-3-030-00794-2_10. URL https://www.researchgate.net/publication/327520269.

Aarne Ranta, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina. Abstract syntax as interlingua: Scaling up the grammatical framework from controlled languages to robust pipelines. *Computational Linguistics*, 46(2):425–486, 2020. doi: 10.1162/coli_a_00378. URL https://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00378.

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.

Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. European Language Grid: An Overview. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3359–3373, Marseille, France, 5 2020a. European Language Resources Association (ELRA).

Georg Rehm, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon. The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3315–3325, Marseille, France, 5 2020b. European Language Resources Association (ELRA).

Signe Rirdance and Andrejs Vasiljevs. *Towards consolidation of European terminology resources: experience and recommendations from EuroTermBank project*. Tilde, 2006.

Laura Rituma, Baiba Saulīte, and Gunta Nešpore-Bērzkalne. Latviešu valodas sintaktiski marķētā korpusa gramatikas modelis. *Language: Meaning and Form*, 10:200–216, 2019. doi: 10.22364/vnf.10.16. URL https://www.apgads.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/Valoda-nozime-forma/VNF-10/vnf_10-16_Nespore_Saulite_Rituma.pdf. The grammar model of Latvian Treebank.

Laura Rituma, Lauma Pretkalniņa, Baiba Saulīte, Gunta Nešpore-Bērzkalne, and Normunds Grūzītis. Latvian Treebank v2.9, 2021. URL http://hdl.handle.net/20.500.12574/56. CLARIN-LV digital library at IMCS, University of Latvia.

Roberts Rozis and Raivis Skadiņš. Tilde MODEL - Multilingual Open Data for EU Languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden, May 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-0235.

Askars Salimbajevs and Indra Ikauniece. System for Speech Transcription and Post-Editing in Microsoft Word. In *INTERSPEECH*, 2017.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Inguna Skadina. Languages of Baltic Countries in Digital Age. In Audrone Lupeikiene, Olegas Vasilecas, and Gintautas Dzemyda, editors, *Databases and Information Systems - 13th International Baltic Conference, DB&IS 2018, Trakai, Lithuania, July 1-4, 2018, Proceedings*, volume 838 of *Communications in Computer and Information Science*, pages 32–40. Springer, 2018. doi: 10.1007/978-3-319-97571-9_5. URL https://doi.org/10.1007/978-3-319-97571-9_5.

Inguna Skadina, Ilze Auzina, Daiga Deksne, Raivis Skadins, Andrejs Vasiljevs, Madara Gailuna, and Ieva Portnaja. Filling the Gaps in Latvian BLARK: Case of the Latvian IT Competence Centre. In Inguna Skadina and Roberts Rozis, editors, *Human Language Technologies - The Baltic Perspective - Proceedings of the Seventh International Conference Baltic HLT 2016, Riga, Latvia, October 6-7, 2016*, volume 289 of *Frontiers in Artificial Intelligence and Applications*, pages 3–11. IOS Press, 2016. doi: 10.3233/978-1-61499-701-6-3. URL https://doi.org/10.3233/978-1-61499-701-6-3.

Raivis Skadiņš and Askars Salimbajevs. The COMPRISE cloud platform. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 108–111, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-64-1. URL https://aclanthology.org/2020.iwltp-1.16.

Raivis Skadins, Marcis Pinnis, Arturs Vasilevskis, Andrejs Vasiljevs, Valters Sics, Roberts Rozis, and Andis Lagzdins. Language Technology Platform for Public Administration. In Utka Andrius, Vaicenoniene Jurgita, Kovalevskaite Jolantai, and Kalinauskaite Danguole, editors, *Human Language Technologies - The Baltic Perspective - Proceedings of the Ninth International Conference Baltic HLT 2020, Kaunas, Lithuania, September 22-23, 2020*, volume 328 of *Frontiers in Artificial Intelligence and Applications*, pages 182–190. IOS Press, 2020. doi: 10.3233/FAIA200621. URL https://doi.org/10.3233/FAIA200621.

Inguna Skadiņa. Datorlingvistika Latvijā. In *Nacionālā Enciklopdija*. Latvijas Nacionala enciklopedija, 2021.

Inguna Skadiņa, Andrejs Veisbergs, Andrejs Vasiļjevs, Tatjana Gornostaja, Iveta Keiša, and Alda Rudzīte. *Latviešu valoda digitālajā laikmetā – The Latvian 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/latvian. Georg Rehm and Hans Uszkoreit (series editors).

Inguna Skadiņa, Ilze Auziņa, Normunds Grūzītis, and Arturs Znotiņš. Clarin in Latvia: From the preparatory phase to the construction phase and operation. In *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries (DHN)*, pages 342–350, 2020. URL http://ceur-ws.org/Vol-2612/short21.pdf.

Andrejs Spektors. Latviešu valodas datorfonda izveide. Number 2, pages 74–82. LZA, 2001.

Andrejs Spektors, Ilze Auzina, Roberts Dargis, Normunds Gruzitis, Peteris Paikens, Lauma Pretkalnina, Laura Rituma, and Baiba Saulite. Tezaurs.lv: the largest open lexical database for Latvian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/1095_Paper.pdf.

Andrejs Spektors, Lauma Pretkalniņa, Normunds Grūzītis, Pēteris Paikens, Laura Rituma, and Baiba Saulīte. Tēzaurs.lv 2020, 2019. URL http://hdl.handle.net/20.500.12574/9. CLARIN-LV digital library at IMCS, University of Latvia.

Jörg Tiedemann. OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia, May 30–June 1 2016. Baltic Journal of Modern Computing. URL https://aclanthology.org/2016.eamt-2.8.

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.

Pēteris Vanags. ”Latviešu valoda”. In *Nacionālā Enciklopdija*. Latvijas Nacionala enciklopedija, 2021.

Andrejs Vasiljevs and Inguna Skadina. Latvian language resources and tools: Assessment, description and sharing. In Arvi Tavast, Kadri Muischnek, and Mare Koit, editors, *Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012, Tartu, Estonia, 4-5 October 2012*, volume 247 of *Frontiers in Artificial Intelligence and Applications*, pages 265–272. IOS Press, 2012. doi: 10.3233/978-1-61499-133-5-265. URL https://doi.org/10.3233/978-1-61499-133-5-265.

Andrejs Vasiļjevs, Jana Ķikāne, and Raivis Skadiņš. Development of HLT for Baltic languages in widely used applications. In *Proceedings of the first Baltic conference” Human Language Technologies-the Baltic Perspective*, 2004.

Rinalds Vīksna and Inguna Skadiņa. Large Language Models for Latvian Named Entity Recognition. In *Human Language Technologies–The Baltic Perspective*, pages 62–69. IOS Press, 2020.

Artūrs Znotiņš. NLP-PIPE: Latvian NLP Tool Pipeline, 2015. URL http://hdl.handle.net/20.500.12574/4. CLARIN-LV digital library at IMCS, University of Latvia.

Arturs Znotins and Guntis Barzdins. LVBERT: Transformer-Based Model for Latvian Language Understanding. In *Human Language Technologies - The Baltic Perspective*, volume 328, pages 111–115. IOS Press, 2020. doi: 10.3233/FAIA200610. URL http://ebooks.iospress.nl/volumearticle/55531.

Arturs Znotins and Elita Cirule. NLP-PIPE: Latvian NLP Tool Pipeline. In *Human Language Technologies - The Baltic Perspective*, volume 307, pages 183–189. IOS Press, 2018. doi: 10.3233/978-1-61499-912-6-183. URL http://ebooks.iospress.nl/volumearticle/50320.

Arturs Znotins, Kaspars Polis, and Roberts Dargis. Media monitoring system for Latvian radio and TV broadcasts. In *INTERSPEECH*, 2015.