EUROPEAN LANGUAGE EQUALITY

D1.19

Report on the Icelandic Language

Author	Eiríkur Rögnvaldsson
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	D1.19
Deliverable title	Report on the Icelandic Language
Type	Report
Number of pages	22
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Author	Eiríkur Rögnvaldsson
Reviewers	Jane Dunne, Sabine Kirchmeier
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

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

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
DLE	Digital Language Equality
DMII	Database of Modern Icelandic Inflection
EGIDS	Expanded Graded Intergenerational Disruption Scale
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRC	European Language Resource Coordination
GPU	Graphics Processing Unit
HPC	High-Performance Computing
IGC	Icelandic Gigaword Corpus
IT	Information Technology
LT	Language Technology/Technologies
LTPI	Language Technology Programme for Icelandic
META-NET	EU Network of Excellence to foster META
NGO	Non-Governmental Organisation
NLG	Natural Language Generation
NLP	Natural Language Processing
R&D	Research and Development
SÍM	Samstarf um íslenska máltækni / Consortium on Icelandic Language Technology

Abstract

This report describes the current situation of Icelandic Language Technology. The digital revolution has had great effects on the environment of the Icelandic language. English is now everywhere – on the Internet, in smartphones, in computer games, in voice-controlled applications, etc. This has put Icelandic under great external pressure and threatens the digital vitality of the language. In order to turn the tables, the Icelandic Government has launched and financed a four-year *Language Technology Programme for Icelandic* (LTPI) which started in 2019.

The SÍM consortium, comprising members from academia, NGOs and the private sector, was formed in order to implement the programme. This consortium has already built and developed many language resources and tools from scratch and enhanced and improved a number of pre-existing resources and tools. Among these are a number of text corpora, both large general purpose corpora and smaller specialised corpora, automatically parsed corpora and large audio corpora, new or improved taggers and parsers and machine translation models.

The LTPI is still ongoing and thus, many of its expected deliverables are not yet finalised. However, prototypes of some of them have been made and look promising. When the programme ends by the end of 2022, the situation for Icelandic with respect to language technology will have improved considerably. However, in spite of this, Icelandic will remain a low-resourced language compared to most official European languages. Thus, it is extremely important that R&D work on Icelandic LT will be maintained beyond the current programme.

Útdráttur

Fyrir 10 árum lét evrópska samstarfsnetið META-NET taka saman skýrslur um stöðu máltækni fyrir 30 evrópsk tungumál – *The META-NET White Papers* (Rehm and Uszkoreit, 2012). Ein þessara skýrslna fjallaði um íslensku, *Íslensk tunga á stafrænni öld / The Icelandic Language in the Digital Age* (Rögnvaldsson et al., 2012). Í þeirri skýrslu kom fram að staða íslensku á þessu sviði væri mjög bágborin. Íslenska var eitt fjögurra tungumála sem lentu í neðsta flokki á þeim fjórum sviðum máltækninnar sem voru borin saman, og reyndist standa næstverst þessara 30 tungumála hvað varðar máltæknistuðning.

Nú, 10 árum síðar, stendur evrópska samstarfsnetið ELE, European Language Equality, fyrir gerð nýrra skýrslna um núverandi stöðu mála. Meira en 40 rannsóknastofnanir og háskólar sem búa yfir sérþekkingu í yfir 30 evrópumálum hafa tekið höndum saman og safnað gífurlegum upplýsingum sem veita góða yfirsýn yfir tæknilegan stuðning við tungumálin. Tilgangurinn er sá að finna hvar skórinn kreppir og hvaða hindranir eru í vegi áframhaldandi þróunar í rannsóknum og tækni í þágu tungumálanna. Þessi yfirsýn er forsenda fyrir ítarlegum áætlunum um það hvernig stafrænni jafnstöðu evrópskra tungumála verði náð árið 2030.

Staða íslensku í stafrænum heimi hefur snarversnað undanfarinn áratug vegna ýmissa tæknibreytinga. Tilkoma snjallsíma veldur því að margt fólk er nú sítengt við enskan menningarheim í gegnum netið. Í stað þess að horfa á íslenskar sjónvarpsstöðvar þar sem allt efni er talsett eða textað á íslensku nýtir ungt fólk nú aðallega efnis- og streymisveitur eins og YouTube og Netflix þar sem mestallt efni er á ensku og framboð íslensks efnis af skornum skammti. Tölvuleikir sem alltaf hafa aðallega verið á ensku eru nú iðulega gagnvirkir og spilaðir á netinu sem kallar á málleg samskipti spilara, oft á ensku. Síðast en ekki síst veldur sprenging í notkun raddstýringar umdæmismissi íslenskunnar þar sem raddstýrð tæki skilja yfirleitt ekki íslensku.

Þessi versnandi staða málsins, ásamt áhyggjum vegna þeirra upplýsinga sem komu fram

í skýrslu META-NET um stöðu íslenskrar máltækni, leiddi til þess að árið 2014 var þingsályktun um að gerð skyldi aðgerðaáætlun um notkun íslensku í stafrænni upplýsingatækni samþykkt einróma á Alþingi. Í framhaldi af því ákvað ríkisstjórnin árið 2017 að ráðast í og fjármagna sérstaka máltækniáætlun til fjögurra ára. Þessi áætlun hófst svo síðla árs 2019.

Sjálfseignarstofnuninni Almannarómi var falin umsjón með áætluninni, en samið var við SÍM-hópinn, Samstarf um íslenska máltækni, um framkvæmd hennar – nauðsynlegt rannsóknar- og þróunarstarf. Þátttakendur í SÍM eru Háskóli Íslands, Háskólinn í Reykjavík, Stofnun Árna Magnússonar í íslenskum fræðum, Ríkisútvarpið, Blindrafélagið, Hljóðbókasafnið, Creditinfo-Fjölmiðlavaktin og þrjú sprotafyrirtæki – Miðeind, Tiro og Grammatek.

SÍM-hópurinn hefur nú unnið að framkvæmd máltækniáætlunar í tvö ár og skilað af sér ýmsum afurðum, bæði gagnasöfnum og hugbúnaði. Sumt af þessu hefur verið byggt upp frá grunni en í öðrum tilvikum hafa eldri gögn og búnaður verið aukin og endurbætt. Segja má að nær öll máltæknigögn og hugbúnaður sem nú eru til fyrir íslensku séu bein eða óbein afurð máltækniáætlunarinnar. Allar afurðir áætlunarinnar eru vistaðar í varðveislusafni CLARIN-IS þar sem þær eru öllum aðgengilegar án endurgjalds, yfirleitt með stöðluðum opnum leyfum. Nokkrar þær helstu eru taldar hér á eftir.

Risamálheildin er safn margvíslegra texta, einkum frá síðustu 20 árum, samtals 1,64 milljarðar orða. Stærstur hluti textanna er úr dagblöðum og vefmiðlum en þar eru einnig Alþingisræður, dómar, fræðslutextar af Vísindavefnum og Wikipediu, og fleira. Textarnir eru málfræðilega greindir og þeim fylgja margvíslegar bókfræðilegar upplýsingar.

Tvær stórar vélþáttaðar málheildir eru til. *GreynirCorpus* hefur að geyma 10 milljónir setninga sem hafa verið stofnhlutagreindar í setningatré. *Samtímalegi íslenski trjábankinn* inniheldur 30 milljónir setninga úr *Risamálheildinni* sem hafa verið þáttaðar með tauganetsþáttara.

Einnig eru til ýmsar smærri málheildir til sérhæfðra nota, svo sem *Íslenska villumálheildin* með textum þar sem margvíslegar villur hafa verið merktar, *Íslenska lesblinduvillumálheildin* með villugreindum textum skrifuðum af lesblindum, og *Íslenska bannorðamálheildin* með óviðeigandi eða gildishlöðnum orðum.

Helsta samhliða málheildin fyrir íslensku er *ParIce* sem hefur að geyma samskipaða texta á íslensku og ensku, alls 3,5 milljón þýðingareininga. Einnig er til risastór bakþýdd þjálfunarmálheild fyrir tauganetsþýðingar (44,7 milljónir setninga úr ensku og 31,3 milljónir úr íslensku).

Fáein hljóðsöfn eru til, einkum *Talrómur*, upptökur af átta málhöfum, samtals 12.780 mínútur, og *Málrómur*, raddsýni frá 563 málhöfum, samtals 9.000 mínútur. Gríðarstórt safn, *Samrómur*, er nú í uppbyggingu með aðferðum hópvirkjunar. Í nóvember voru komin inn raddsýni frá 22.000 málhöfum, samtals 135.000 mínútur.

Beygingarlýsing íslensks nútímamáls hefur verið í þróun undanfarin 20 ár en hefur verið uppfærð innan máltækniáætlunarinnar. Hún hefur að geyma um 305 þúsund uppflettiorð og rúmlega sex milljónir beygingarmynda. *BÍN-kjarni* hefur að geyma beygðan grunnorðaforða málsins, um 58 þúsund orð.

Önnur mikilvæg orðasöfn eru *Íslensk framburðarorðabók* með 59 þúsund hljóðrituðum orðmyndum og *Íslensk orðskiptingaskrá* með 218 þúsund orðum þar sem möguleg línuskipting er sýnd. *Íslensk nútímamálsorðabók* með 56 þúsund orðum er einnig í varðveislusafni CLARIN-IS en hefur ekki verið unnin innan máltækniáætlunarinnar.

Ýmiss konar hugbúnaður til málfræðilegrar greiningar hefur verið þróaður. Þar má helst nefna tvo hugbúnaðarvöndla sem hvor um sig inniheldur ýmis forrit. *IceNLP* var upphaflega gerður á árunum 2005–2008 og inniheldur tilreiðara, markara, lemmald og grunnþáttara. *Greynir* er nýrri vöndull sem þáttar texta, greinir lemmur, beygir nafnliði, greinir í orðflokka o.fl.

Auk þessa má nefna *ABL Tagger* sem nær 96,95% nákvæmni í mörkun texta og meðfylgjandi lemmald sem nær 98,3% nákvæmni í lemmun. Einnig hafa tveir tauganetsþáttarar

verið þróaðir; *Tauganetsþáttari Miðeindar* og *Íslenska tauganetsþáttunarpípan*. Báðir byggjast á Berkeley tauganetsþáttaranum.

Fyrir 10 árum þróaði Google talgreiningu fyrir íslensku, í samvinnu við íslenska vísindamenn. Um svipað leyti gerði pólska fyrirtækið Ivona, sem nú er í eigu Amazon, íslenskan talgervil fyrir Blindrafélagið. Þessi búnaður hefur nýst vel en þarfnast endurnýjunar og er auk þess ekki opinn. Ýmis búnaður til talgreiningar og talgervingar er í smíðum innan máltækniáætlunarinnar en er ekki enn kominn á markað.

Vélþýðingar eru eitt af áherslusviðum máltækniáætlunarinnar og ýmis stoðtól til vélþýðinga hafa verið útbúin. Miðeind hefur unnið að þróun tauganetsþýðinga milli íslensku og ensku sem lofa mjög góðu. Opnuð hefur verið vefsíða þar sem fólk getur látið kerfið þýða texta.

Í *Stefnu Íslands um gervigreind* sem var gefin út í apríl 2021 er sérstaklega tekið fram að til þess að íslenska standi jafnfætis öðrum tungumálum í heiminum sé nauðsynlegt að þróa innviði sem tryggi að hún verði nothæf í heimi tækninnar. Í stjórnarsáttmála nýrrar ríkisstjórnar sem tók við völdum í lok nóvember 2021 segir að áfram verði unnið að því að styrkja stöðu íslenskunnar í stafrænum heimi með áherslu á máltækni, og markáætlun um samfélagslegar áskoranir, m.a. á sviði máltækni, verði haldið áfram allt kjörtímabilið.

Máltækniáætlun stjórnvalda hefur verið mikil lyftistöng fyrir íslenska máltækni. Fyrir utan það að byggja upp gögn og þróa hugbúnað eins og lýst er hér að framan hefur máltækniáætlunin leitt saman háskóla, rannsóknastofnanir, félagasamtök og fyrirtæki sem hafa átt mjög frjóa og árangursríka samvinnu. Fjöldi rannsakenda og stúdenta sem vinna að máltækniverkefnum hefur margfaldast, og stúdentum í máltækninámi fjölgað að mun. Það er gífurlega mikilvægt að þessu starfi verði haldið áfram. Vinnu að því að gera íslensku jafnsetta öðrum tungumálum í stafrænum heimi lýkur aldrei.

1 Introduction

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

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

The report has been developed in the frame of the European Language Equality (ELE) project. 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.

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they can be discovered, browsed and further investigated by means of comparative visualisations across languages.

2 The Icelandic Language in the Digital Age

2.1 General Facts

Icelandic is a North Germanic language with its roots in Old Norse. It is the only official language of Iceland apart from Icelandic Sign Language. Even though it is only spoken in Iceland and only by around 370,000 people, it is not considered endangered according to UNESCO's Language Vitality Scales[2] or EGIDS.[3] Icelandic has until very recently been the first language of virtually all inhabitants. The language community is very homogeneous, and dialectal variation is negligible. Icelanders are known for their language purism and during the past thousand years, Icelandic has changed less than related languages, although the changes are more extensive than commonly believed.

Icelandic is a morphologically rich language. Nouns and pronouns inflect for case and number, and can have one of three genders. Adjectives inflect for case, number, gender, definiteness, and grade. Verbs are conjugated for person, number, mood, tense and voice. The language is fusional, such that a single ending usually stands for more than one morphological category. The inflectional system is further complicated by a great number of inflectional and conjugational classes, such that the same morphological category, or combination of categories, is represented by a number of different endings depending on the stem.

Typologically, Icelandic is a SVO (subject-verb-object) language with a strong V2 rule that requires the verb to appear in the second (or first) position of the sentence. However, because of the rich inflectional system, word order is relatively free; certain words can be moved around without the meaning of the sentence being altered or lost. The large number of inflectional forms, the free word order, and productive word formation processes make morphosyntactic tagging quite a challenge. The most widely used Icelandic tagset contains around 700 different morphosyntactic tags, but a simplified version has recently been developed.

The Icelandic alphabet is based on the Latin alphabet with a number of additions, especially vowel symbols with an acute accent, *á é í ó ú ý Á É Í Ó Ú Ý*, and the vowel symbols *æ Æ* and *ö Ö* which are also used in a number of other languages. Furthermore, Icelandic employs two more eccentric symbols – *ð Ð* (eth, not to be confused with "d with a stroke", *đ*) which is also used in Faroese, and *þ Þ* (thorn) which is not used in any other language.

A few years ago, it could be maintained that Icelandic was used – and was in fact the only language used – in all spheres of society: in government and administration; in education; in business and commerce; in the mass media; in cultural life; on the Internet; and in ordinary face-to-face communication. According to all vitality scales, the language should therefore be safe, but in the last decade or so, Iceland has gone through dramatic societal and technological changes which have led to a massive increase in the use of English in the Icelandic language community and thus an increase in the external pressure on the language.

In the last decade, Iceland had an explosion in tourism. As a result, advertisements, signs, menus etc. are often directed towards tourists and thus only in English. A number of cultural events are also introduced and performed in English to attract tourists. The number of foreign workers has grown considerably – people of foreign origin now amount to more than 15% of the population and many of them work in restaurants or shops where they have to communicate with customers, usually in English. English is also increasingly being used in higher education – more and more university courses are taught in English. Furthermore, ongoing globalisation might affect young people's attitudes towards Icelandic – they want to study abroad, work abroad and live abroad and know that Icelandic is of little use outside

[2] https://ich.unesco.org/doc/src/00120-EN.pdf
[3] https://www.ethnosproject.org/expanded-graded-intergenerational-disruption-scale/

Iceland.

2.2 Icelandic in the Digital Sphere

Iceland has the highest percentage of Internet users in Europe. In 2020, 98% of Icelandic households had Internet access.[4] In the same year, 68,344 websites had .is as the top level domain.[5] The Internet, smartphones and tablets have revolutionised the daily lives of people, especially children and teenagers who are now online 24/7, so to speak. They are directly connected to the digital world which is for the most part in English. Icelandic is sufficiently represented on the Internet, with a number of media websites and an Icelandic Wikipedia, for instance, but most people also frequently visit news sites in English, access various types of information in English, etc. Even though Icelandic is the main language used on social media, English is also prominent.

Research has shown that the majority of children and young people no longer watch old-fashioned linear TV but watch material from service and content providers like Netflix and YouTube instead. This is important since all programs on Icelandic TV are in Icelandic – either dubbed, as all programs intended for young children, or with Icelandic subtitles. Netflix and YouTube, on the other hand, offer very limited material in Icelandic, although the situation has improved somewhat in the past two years.

Computer games, which are especially played by young people, are overwhelmingly in English. They are becoming more and more interactive, which means that players are not only reacting to actions in the game, as used to be the case, but communicating – either with the game itself, or with other players. Since these players may be spread around the globe, the language of communication is often English.

The technological change that might have the most far-reaching consequences for Icelandic, is the ongoing explosion in the use of voice control. A few years ago, Icelandic authorities started to realise what this would entail for speakers of a language like Icelandic, which has up to now not been usable within this new technology. The main reason for establishing the LTPI, which started in 2019, was to make Icelandic usable in the digital sphere and both speech recognition and speech synthesis are among the core areas of the programme.

3 What is Language Technology?

Natural language[6] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

The computational processing of human languages has been established as a specialized field known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP) or, more generally, Language Technology (LT). While there are differences in focus and orientation, since CL is more informed by linguistics and NLP by computer science, LT is a more neutral term. In fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably AI), mathematics and psychology among others. In practice, these commu-

[4] https://www.statista.com/statistics/185663/internet-usage-at-home-european-countries/
[5] https://www.isnic.is/is/tolur
[6] This section has been provided by the editors. It is an adapted summary of Agerri et al. (2021) and of Sections 1 and 2 of Aldabe et al. (2021).

nities work closely together, combining methods and approaches inspired by both, together making up *language-centric AI*.

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

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

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

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

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

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

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

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

4 Language Technology for Icelandic

The Icelandic Government launched the LTPI in September 2019 (Nikulásdóttir et al., 2017). The resources and tools built within this programme are available for free under standard open licenses. Most of the existing resources and tools for Icelandic are direct or indirect outputs of this programme (Nikulásdóttir et al., 2020). Many of them have been built from scratch, but in other cases, existing resources and tools have been updated and enhanced. Thus, they are all up to date. A number of the most important language resources and tools are briefly described below, but a more detailed description of most of them can be found in (Nikulásdóttir et al., 2022). Almost all of these resources and tools are stored in the CLARIN-IS repository.[8]

4.1 Language Data

Monolingual Text Corpora

The Icelandic Gigaword Corpus (IGC) is a monolingual corpus comprising almost 1.9 billion tokens. Most of the texts are from 2001-2020. They represent different genres, although the majority consists of newspaper articles and transcribed radio and television news. Other important genres are social media texts, parliamentary speeches and adjudications. The Icelandic Wikipedia is also included, in addition to a number of smaller genres. However, transcribed spoken language texts are severely underrepresented. The corpus is morphosyntactically tagged and contains various information on the source texts. A number of subcorpora

[7] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at an annual rate of 18.4% from 2020 to 2028 (https://www.globenewswire.com/news-release/2021/03/22/2196622/0/en/Global-Natural-Language-Processing-Market-to-Grow-at-a-CAGR-of-18-4-from-2020-to-2028.html). A different report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at USD 13 billion in the year 2020 and is projected to reach USD 25.7 billion by 2027, growing at an annual rate of 10.3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

[8] https://repository.clarin.is

have also been made available separately (adjudications, books, journals, laws, parliamentary speeches, social media).

A few parsed corpora exist, with most of them having been automatically parsed. *Greynir-Corpus* contains 10 million sentences from news sources which have been parsed into full constituency trees. It is accompanied by a gold standard corpus of 5,000 manually parsed sentences. *The Icelandic Contemporary Corpus* is a constituency parsed corpus built by using an Icelandic model of the Berkeley Neural Parser and containing 30 million clauses from the IGC.

Even though some of the above-mentioned corpora are fairly large, they are far from being large enough to develop and train AI models. Furthermore, some important genres are missing. Thus, spoken language is only a small fragment of the IGC and not present at all in the other corpora. It is very important to build corpora including these genres but it is expensive and accessing relevant data is difficult, not least because of GDPR issues.

In addition to these general purpose corpora, a number of small specialised corpora have been developed, such as the *Icelandic Error Corpus* which is a collection of texts in modern Icelandic annotated for mistakes related to spelling, grammar, and other issues (4,046 texts with 56,956 error instances classified into 253 categories); the *Icelandic Dyslexia Error Corpus* (26 texts with 5,730 categorized error instances); and the *Icelandic Taboo Database* which is a list of words that may in some way be considered inappropriate, taboo and/or loaded in use or meaning (2,724 words).

Bi- and Multilingual Text Corpora

There exists a number of bilingual English-Icelandic corpora. Most of them are domain-specific corpora from ELRC and are not aligned. Furthermore, they are rather small, with the exception of corpora from the *ParaCrawl Project*.[9] However, a few general purpose aligned corpora exist, the most important being *ParIce* with 3.5 million translation units. There is also a synthetic back-translated training corpus for neural machine translation containing 76 million translation units. It is evident that much larger bilingual corpora are needed, especially between Icelandic and English but also between Icelandic and other languages such as Icelandic and Polish.

Multimodal Corpora

A few audio corpora exist. The most important one is *Talrómur* which consists of 122,417 short audio clips of eight different speakers reading short sentences – 12,780 minutes in total. Another is *Málrómur* which contains voice samples from 563 individuals, around 9,000 minutes. However, a large crowdsourcing project, *Samrómur*, is now ongoing. In November, a total of 1.55 million sentences from 22,000 speakers had been recorded, 135,000 minutes in all. At the end of the project, all the recordings will be made available for free as other deliverables of the LTPI.

No video corpora have been built for Icelandic.

Lexical/Conceptual Resources

The Database of Modern Icelandic Inflection (DMII) is supposed to contain the inflectional paradigms of the whole vocabulary of Icelandic. The development of this resource started in 2002, and it has contributed to most language resources and tools that have been developed for Icelandic, either directly or indirectly. The current version has a vocabulary of approx.

[9] https://paracrawl.eu

305,000 lemmas, 6.2 million inflectional forms. The *DMII Core* is an extract of DMII and contains the core vocabulary of Modern Icelandic, around 58,000 entries.

The Dictionary of Contemporary Icelandic is a monolingual dictionary with 56,000 entries which is constantly being updated. Sound files with recordings of all the headwords in the dictionary are also available. Other important resources are *Icelandic Pronunciation Dictionary* with 59,000 entries, *Icelandic Hyphenation Dictionary* with 218,000 entries, and *Icelandic Term Bank* containing 104,000 entries from 41 different term collections covering widely different fields.

Models and Grammars

The company Miðeind, which is a member of the SÍM Consortium, is developing AI models for machine translation and some of them are already available, such as *GreynirTranslate – mBART25 NMT*, general domain IS-EN and EN-IS translation models based on a multilingual BART model. *Icegrams* is a package that encapsulates a large trigram library for Icelandic (14 million unique trigrams and their frequency counts). However, much larger and better models are clearly needed for developing various LT applications.

4.2 Language Technologies and Tools

Text Analysis

There exist a number of tools for analysing Icelandic text. Among them are two packages that each include various tools. *IceNLP* is a package which was originally developed between 2005-2008 and contains a tokeniser, part-of-speech tagger, a lemmatiser, and a shallow parser. Some of these components have recently been updated. *Greynir* is a more recent package that can parse text into constituency trees, find lemmas, inflect noun phrases, assign part-of-speech tags and more. It uses a tokeniser by the same authors.

ABL Tagger is a part-of-speech tagger that achieves an accuracy of 96.95% using the MIM-Gold tagset. It is accompanied by a lemmatiser which achieves an accuracy of 98.3%. *The Icelandic Neural Parsing Pipeline* includes all steps necessary for parsing plain Icelandic text. The preprocessing step consists of tokenization, both punctuation and matrix clause splitting. The parsing step consists of an Icelandic model of the Berkeley Neural Parser which reports an 84.74 F1 score. The *Miðeind Neural Constituency Parser* is an experimental variant of the Berkeley Neural Parser architecture.

Speech Processing

Ten years ago, Google developed speech recognition for Icelandic in cooperation with Icelandic researchers. Around the same time, a speech synthesiser for Icelandic was developed by the Polish company Ivona which is now a subsidiary of Amazon. Although these applications have been very useful, they are now outdated and furthermore privately owned which severely limits their use. A number of tools for speech processing are currently being developed within the LTPI, among them both a new speech recogniser and a speech synthesiser, but these tools are not yet publicly available although prototypes of them have been publicly demonstrated.

Embla is the first voice assistant app for the Icelandic language, available both for iPhone and Android smartphones. It combines a speech recogniser, a speech synthesiser and the *Greynir* tool which it uses to search for answers to questions that the user poses. When the answer is found it is formulated in the correct Icelandic phrase taking into account inflection and other grammatical features. Finally a fully-formatted response is sent to the synthesiser.

Translation Technologies

Machine translation is one of the core areas in the LTPI. Miðeind has been developing a translation system between English and Icelandic using neural networks. Although this system is still under development, it already gives very promising results. The pilot version is offered as a web-based service.[10]

Information Extraction and Information Retrieval

Greynir extracts information from Icelandic text which allows natural language querying of that information and facilitates natural language understanding. *Greynir* periodically scrapes chunks of text from Icelandic news sites on the web. The text is tokenised, tagged and parsed according to a hand-written context-free grammar for Icelandic. The resulting parse trees are then stored in a database and processed by grammatical pattern matching modules to obtain statements of fact and relations between stated facts.

Language Generation and Summarisation

Tools for Language Generation and Summarisation are lacking.

Human-Computer Interaction

With the exception of the *Embla* voice assistant app described above, there is a lack of tools for Human-Computer Interaction.

All deliverables of the LTPI will be deposited to the CLARIN-IS repository.[11] They can be downloaded from the repository for free, most of them under standard open licenses, and used in any kind of application. Since the LTPI will continue until the end of 2022, a number of resources and tools will be built and developed in the next months. Some of them already exist in demo or prototype versions but have not been made publicly available.

Most of the deliverables of the programme up to now have been basic language resources and tools, such as text and audio corpora and various tools for text analysis. A few such resources and tools existed previously but have been greatly enhanced or replaced by new and better ones. Advanced applications built on these resources and tools, such as speech recognisers, speech synthesisers, spell and grammar checkers and machine translation systems are under development within the programme. Prototypes or demo versions of most of them have already been made and are offered as web-based applications. There is no reason to doubt that mature versions of these applications will be available by the end of the programme.

4.3 Projects, Initiatives, Stakeholders

The Icelandic Government published an AI strategy document in April 2021.[12] The document describes three pillars on which the AI policy for Iceland shall rest. The first is that AI should be for the benefit of everyone. The report points out that there are many potential situations in which decisions made by AI systems may have ethical and moral implications, and suggests guiding values along with a framework for evaluating such circumstances. The importance of developing LT resources and tools for Icelandic is explicitly mentioned.

[10] https://velthyding.is
[11] https://repository.clarin.is
[12] https://www.stjornarradid.is/library/01--Frettatengt---myndir-og-skrar/FOR/Fylgiskjol-i-frett/StefnaÍslandsumgervigreind

To ensure the competitiveness of Iceland's private sector, which is the policy's second pillar, the report suggests methods for supporting and incentivising increased digitisation of industry. AI technologies are capable of enabling solutions to complex problems that have previously been uneconomical to solve using manpower in countries with a low population. The third pillar is education. A future of continuous education, local expertise in AI, and opportunities for adapting AI technologies to Iceland's industrial needs, needs to be ensured.

As mentioned above, the five year Government-funded LTPI started in 2019. The estimated total cost of the Programme is around 20 million Euros. The aim of this project is both to build basic language resources and tools and to develop a number of practical applications. Emphasis is laid on five core areas: Speech synthesis, speech recognition, spell and grammar checking, machine translation, and language resources. Furthermore, a strategic research and development programme for language and technology has been established and LT education has been strengthened.

In the policy statement of the new Government that took office in November 2021,[13] it is explicitly stated that the Government will continue supporting the development of Icelandic LT after the LTPI expires and the strategic R&D programme will be prolonged throughout the current election period, until 2025.

The self-owned foundation *Almannarómur*[14] ('voice of the people') was entrusted with the role of conducting the five above-mentioned core tasks. Almannarómur was founded in 2014 with the purpose of developing LT solutions for Icelandic. The initiative came from people in academia who had been working on LT but wanted to get more people involved, and especially to reach out to other sectors of the society.

The founding members were over 20 – universities and research institutions, IT companies, financial institutions, insurance companies, energy companies, companies in the travel industry, and organizations of disabled people. The main emphasis was on raising awareness among companies, politicians and the general public on the opportunities of LT and the importance of LT for the future of the Icelandic language.

Almannarómur, in turn, commissioned the *SÍM Consortium*[15] to carry out the research and development work necessary for this project. The SÍM Consortium consists of two universities, University of Iceland and Reykjavik University, the Árni Magnússon Institute for Icelandic Studies, the National Broadcasting Service, the Association of the Visually Impaired, the Icelandic Audio Library, one established private company and three startup IT companies.

The SÍM Consortium comprises practically all institutions, companies, and people that have been active within LT in Iceland for the past two decades – researchers, developers and language LT users are well represented in the Consortium.

One of the SÍM members, the University of Iceland, participated in the EU-funded PRINCIPLE project from 2019-2021.[16] The main aim of the project was to identify, collect and process high-quality Language Resources for four under-resourced languages (Icelandic, Croatian, Irish and Norwegian). Furthermore, the University of Iceland participates in the ELRC and has collected bilingual data from various public organisations.

5 Cross-Language Comparison

The LT field[17] 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 consid-

[13] https://www.stjornarradid.is/library/05-Rikisstjorn/Agreement2021.pdf
[14] https://almannaromur.is/en
[15] https://icelandic-lt.gitlab.io
[16] https://principleproject.eu
[17] This section has been provided by the editors.

erable increase in the number and quality of resources for many languages have yielded results unforeseeable before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[18] 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

[18] 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.

2. **Fragmentary support**: the language is present in ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[19]

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[20] 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.[21]

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.

[19] 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.

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

[21] Interactive comparison visualisations of the technology support of Europe's languages will be possible on the ELG website using a dedicated dashboard, which dynamically analyses the resources available in the ELG repository, from the middle of 2022 onwards.

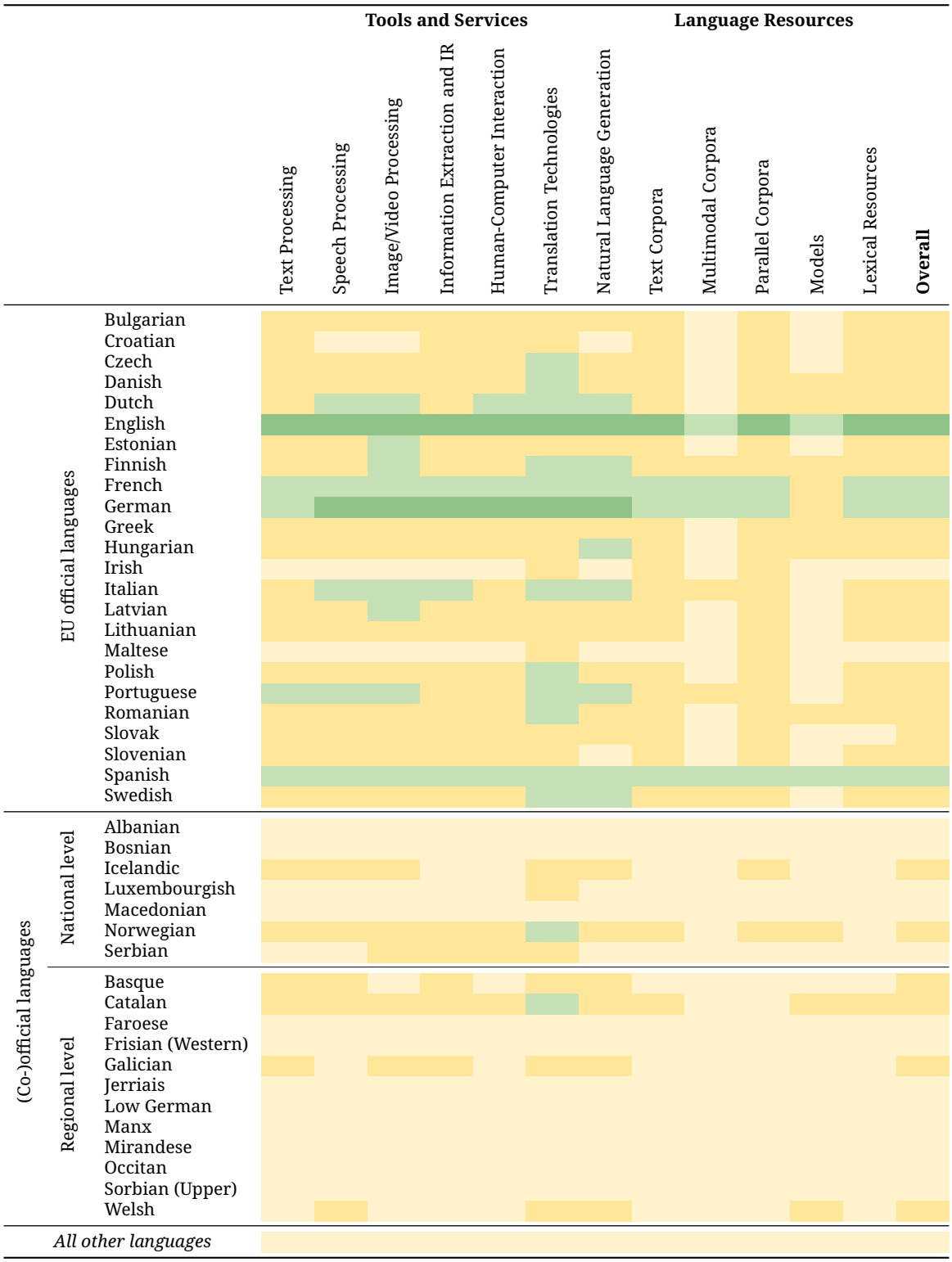

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g., German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[22] 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.

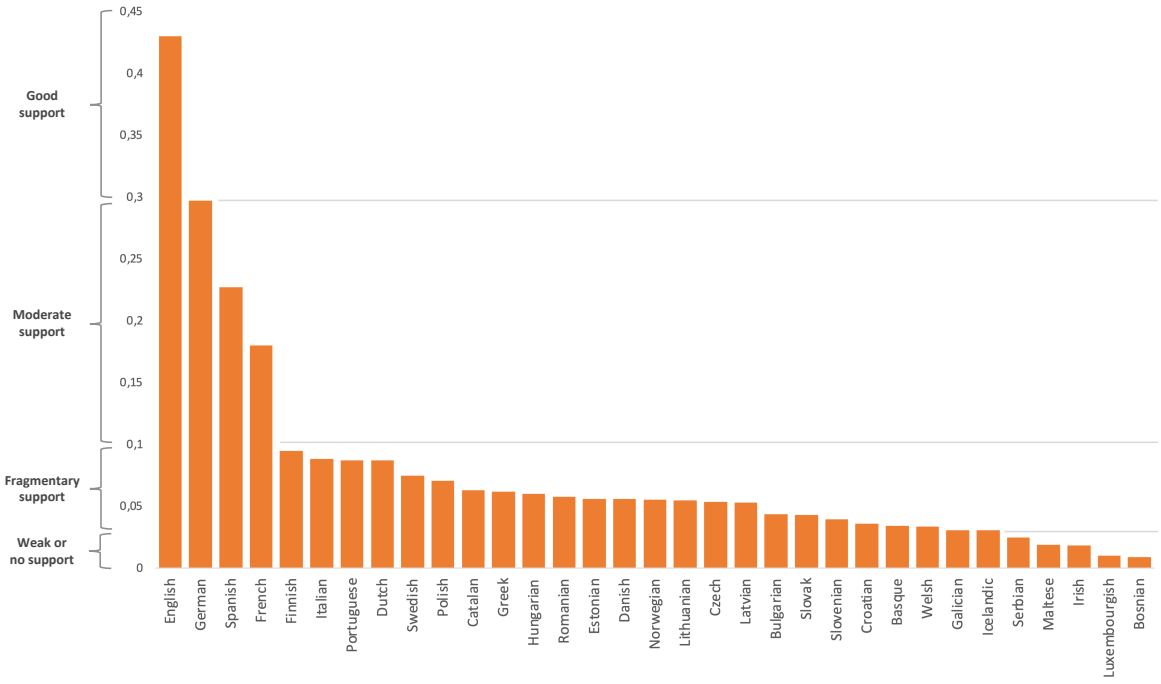

Figure 1: Overall state of technology support for selected European languages (2022)

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,

[22] 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.

emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural language Understanding has been reached.

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

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

Ten years ago, the *META-NET White Paper* on Icelandic highlighted the alarming status of Icelandic LT. Icelandic was one of only four languages receiving the lowest score in all four categories that were evaluated. At that time, funding for R&D in Icelandic LT was nonexistent, no companies were developing or offering LT products, and the digital future of Icelandic didn't look bright. The results of the White Paper raised concerns among politicians and the public and were discussed in the Icelandic Parliament in 2012. In 2014, the Parliament unanimously adopted a resolution on the necessity of making Icelandic usable in the digital domain. This eventually resulted in the implementation and financing of the LTPI.

The LTPI has revolutionised the situation in Icelandic LT. The forming of the SÍM Consortium, which was an indirect result of the programme, has led to a very fruitful cooperation among all stakeholders. Researchers who used to work individually on small projects now work together on implementing projects on a much bigger scale. The number of researchers and students involved in LT has multiplied and new startup companies have grown out of the programme. As described above, many important resources and tools have been built and developed in the two years since the programme started. However, there is still a long way to go. It is to be hoped that the LTPI will deliver high-quality applications that will be welcomed by the public and contribute to the digital vitality of Icelandic.

But even if they do, Icelandic will still be lagging behind the larger European languages. When the LTPI ends, Icelandic will still lack a number of important resources such as spoken language corpora; parallel corpora (Icelandic and other languages than English, such as Polish and the Scandinavian languages); corpora for different purposes (sentiment analysis, question answering, summarisation); annotated multimodal coprora; and term lists.

Furthermore, Icelandic will lack tools for sentiment analysis; summarisation; question answering; natural language understanding; natural language generation; dialogue management; disambiguation; translation between Icelandic and other languages than English; speech translation; automatic subtitling; specialised speech recognition (child language, nonnative Icelandic, real-time subtitling); advanced speech synthesis (intonation, empathy); spe-

cialised speech synthesis (children's voices); specialised grammar checking (for teaching, dyslexic people, non-native speakers) – and a number of other resources and tools.

Hopefully, the Government will keep its promise to support the continuation of the LTPI, but a large-scale European cooperation would be a very welcome assistance in filling these gaps.

References

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

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

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

Anna Björk Nikulásdóttir, Jón Guðnason, and Steinþór Steingrímsson. *Language Technology for Icelandic 2018–2022. Project Plan.* Icelandic Ministry of Education, Science and Culture, 2017. URL https://clarin.is/media/uploads/mlt-en.pdf.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. Language Technology Programme for Icelandic 2019–2023. In Nicoletta Calzolari (Conference Chair), 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 3414—3422, Marseille, France, May 2020. European Language Resources Association (ELRA).

Anna Björk Nikulásdóttir, Þórunn Arnardóttir, Starkaður Barkarson, Jón Guðnason, Þorsteinn Daði Gunnarsson, Gunnar Thor Örnólfsson, Anton Karl Ingason, Haukur Páll Jónsson, Hrafn Loftsson, Hulda Óladóttir, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, Atli Þór Sigurgeirsson, Vésteinn Snæbjarnarson, and Steinþór Steingrímsson. Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS. In Monica Monachini and Maria Eskevich, editors, *Selected Papers from the CLARIN Annual Conference 2021, Virtual Event, 2021, 27–29 September*, Linköping, Sweden, 2022. Linköping Electronic Conference Proceedings.

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.

Eiríkur Rögnvaldsson, Kristín M. Jóhannsdóttir, Sigrún Helgadóttir, and Steinþór Steingrímsson. *Íslensk tunga á stafrænni öld – The Icelandic 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/icelandic. Georg Rehm and Hans Uszkoreit (series editors).

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.