EUROPEAN LANGUAGE EQUALITY

D1.20

Report on the Irish Language

Author	Teresa Lynn
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	D1.20
Deliverable title	Report on the Irish Language
Type	Report
Number of pages	30
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	Teresa Lynn
Reviewers	Maria Giagkou, Victoria Arranz
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

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

Contents

1 Introduction **2**

2 The Irish Language in the Digital Age **3**
2.1 General Facts 3
2.2 Irish in the Digital Sphere 5

3 What is Language Technology? **6**

4 Language Technology for Irish **8**
4.1 Language Data 8
4.2 Language Technologies and Tools 11
4.3 Projects, Initiatives, Stakeholders 12

5 Cross-Language Comparison **14**
5.1 Dimensions and Types of Resources 14
5.2 Levels of Technology Support 15
5.3 European Language Grid as Ground Truth 15
5.4 Results and Findings 16

6 Summary and Conclusions **18**

List of Figures

1 Overall state of technology support for selected European languages (2022) . . 18

List of Tables

1 State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support) 17

List of Acronyms

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
CALL	Computer Assisted Language Learning
CEF	Connecting Europe Facility
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
DL	Deep Learning
DLE	Digital Language Equality
DCU	Dublin City University
DGT	Directorate-General for Translation
DTCAGSM	Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media
EC	European Commission
ELE	European Language Equality *(this project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRC	European Language Resource Coordination
ELRI	European Language Resource Infrastructure
EMEA	Europe, the Middle East, and Africa
EU	European Union
GPU	Graphics Processing Unit
HPC	High-Performance Computing
ICC	International Comparable Corpus
LR	Language Resources/Resources
LSG	Líonra Séimeantach na Gaeilge
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MMID	Massively Multilingual Image Dataset
MT	Machine Translation
MWE	Multi-Word Expression
NCC	National Competence Centre
NCI	National Corpus of Ireland
NCII	New Corpus for Ireland – Irish
NMT	Neural Machine Translation
NLG	Natural Language Generation

NLP	Natural Language Processing
NUIG	National University of Ireland Galway
POS	Part-Of-Speech
PII	Personal Identifiable Information
R&D	Research and Development
RTÉ	Raidió Teilifís Éireann
SMT	Statistical Machine Translation
SR	Speaker Recognition
TCD	Trinity College Dublin
UD	Universal Dependencies
UNESCO	United Nations Educational, Scientific and Cultural Organization
VSO	Verb-Subject-Object

Abstract

In today's digitally connected world, advances in AI lie behind many of the ways in which we work, do business, shop, study and socialise. Language technology underpins many of the applications and platforms that enable our digitally enhanced lives (virtual assistants, search engines, translation tools, spell-checkers, language learning tools, etc.). Yet these advances do not benefit all Irish citizens equally. Due to a lack of sufficient Irish language technologies, Irish speakers often need to revert to using English. Such a language shift plays a major role in the risk of digital extinction, i. e. an eventual decline in language use due to lack of technological support. As such, Irish is in a precarious position while it competes alongside the most technologically supported language in the world. The lack of awareness around the need for immediate strategic planning, investment and development of Irish language technologies further widens the gap. It becomes clear therefore, that attempts to increase language use through increasing the number of speakers alone will be futile if those speakers will need to live in a separate unconnected world. This document sets out to highlight the work carried out in this area thus far, and the gaps and challenges that need to be addressed for this official national and EU language.

We provide an up-to-date overview of the current status of Irish language technology through an analysis of data resources and tools/services listed in the European Language Grid (ELG) – a catalogue that has been populated by language informants across all European member states. The summary shows that only a few areas of Irish speech and language technologies have been addressed (though not fully); mostly through short term funding grants within university settings. In the context of cross-lingual comparisons, we show that relatively minimal changes have happened over the past 10 years to address these gaps. Finally, we demonstrate how the lack of language technology resources, training and education programmes, dedicated funding programmes or strategies, industry collaborations, lack of awareness and mere lack of value assigned to language technology are all contributing factors to Irish being one of the most under-supported languages in Europe.

Achoimre

I ndomhan an lae inniu, atá nasctha go digiteach, is iad forbairtí san Intleacht Shaorga atá taobh thiar de go leor de na bealaí a bhíonn muid ag obair, a ndéanann muid gnó, a bhíonn muid ag siopadóireacht, ag staidéar agus ag bualadh le chéile. Tá an teicneolaíocht teanga mar bhonn faoi chuid mhór de na feidhmchláir agus ardáin a chumasaíonn ár saol, atá feabhsaithe go digiteach (cúntóirí IS, innill chuardaigh, uirlisí aistriúcháin, seiceálaithe litrithe, uirlisí foghlama teanga, etc.) Ní théann na forbairtí sin chun tairbhe ar an mbealach céanna do shaoránaigh uile na hÉireann, áfach. De cheal teicneolaíochtaí teanga leordhóthanacha Gaeilge, go minic bíonn ar chainteoirí Gaeilge filleadh ar an mBéarla. Is mór an ról atá ag aistriú teanga den sórt sin sa riosca a bhaineann le díobhadh teanga, i. e. laghdú ar úsáid teanga i ngeall ar easpa tacaíochta teicneolaíochta. Mar sin, is contúirteach an staid ina bhfuil an Ghaeilge agus í in iomaíocht le cuid de na teangacha is mó tacaíochta teicneolaíochta ar domhan. An easpa feasachta atá ann go bhfuil gá le pleanáil, infheistíocht agus forbairt straitéiseach láithreach i dtaca le teicneolaíochtaí teanga Gaeilge, cuireann sí leis an mbearna sin. Is léir, mar sin, nach fiú iarrachtaí a dhéanamh úsáid na teanga a mhéadú tríd an líon cainteoirí a mhéadú amháin má bhíonn ar na cainteoirí sin maireachtáil i ndomhan ar leithligh nach bhfuil nasctha go digiteach. Sa cháipéis seo féachtar le haird a tharraingt ar an obair atá déanta sa réimse sin go dtí seo, chomh maith leis na bearnaí agus na dúshláin ar gá aghaidh a thabhairt orthu ar mhaithe leis an teanga seo, ar teanga oifigiúil náisiúnta agus teanga oifigiúil de chuid an Aontais í.

Tugaimid léargas cothrom le dáta ar stádas theicneolaíochtaí teanga na Gaeilge le hanailís ar acmhainní sonraí agus uirlisí/seirbhísí atá liostaithe i nGreille Teangacha na hEorpa – catalóg a bhfuil eolas curtha léi ag faisnéiseoirí teanga ar fud Bhallstáit na hEorpa. Léirítear san achoimre nach bhfuil aghaidh tugtha ach ar líon beag réimsí de theicneolaíochtaí urlabhra agus teanga na Gaeilge, go príomha trí dheontais ghearrthéarmacha d'ollscoileanna, agus nach bhfuil sin féin déanta ach go pointe. I gcomhthéacs comparáidí tras-teanga, léirímid gur beag athrú atá déanta le deich mbliana anuas chun dul i ngleic leis na bearnaí sin. Agus ar deireadh, taispeántar gur fachtóirí iad an easpa acmhainní teicneolaíochta teanga, clár oiliúna agus oideachais, clár maoiniúcháin tiomnaithe nó straitéisí tiomnaithe, comhair sna tionscail, easpa feasachta agus an easpa luacha a chuirtear ar an teicneolaíocht teanga, fachtóirí a chuireann le staid na Gaeilge mar cheann de na teangacha is lú tacaíochta san Eoraip.

1 Introduction

This study is part of a series that reports on the results of an investigation into 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 provides a detailed, empirical and dynamic map of technology support for our languages.[1]

The report has been developed by the European Language Equality (ELE) project.[2] With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initiatives, the ELE project develops a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

This study focuses on Irish – an official EU language, and considered a low-resourced language in terms of digital support. 'Low-resourced' not only means that there is a severe lack of speech and language applications available for Irish speakers to use, but it also means that the fundamental tools and language resources required to build these technologies are also lacking.

A number of factors have contributed to this. The Irish context differs greatly from that of other official European languages. While Ireland itself has a strong economy and provides an ideal context in which technological investment and advances can thrive, the sociolinguistic landscape of the country does not lend itself to equality across the country's co-official languages (Irish and English). As a majority English-speaking population, Irish citizens in general can benefit from market-driven advances in English speech and language technologies. However, the smaller Irish-speaking market misses out on such technological investment and development in their own language. As all Irish speaking citizens speak and understand English, bilingual technology users have simply adapted to using English when digital needs arise (e. g. speech-recognition, spell-checking, online searches, etc.). This language shift is a matter of concern and plays a key role in the risk of digital extinction.

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

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

That is to say that, while a language continues to be under-supported through technology, it becomes less relevant in daily digital life and subsequently becomes less spoken.

This report highlights the immediate need for an increase in awareness, focus and action relating to the digital readiness of the Irish language. It presents a broad discussion on the status of the language, its current status with respect to technological support, and the challenges and gaps faced in terms of achieving digital language equality. A cross-comparison is made with other EU languages in terms of digital readiness, and finally some recommendations are made with respect to the first steps that can be taken towards securing a place for the Irish in the digital sphere.

2 The Irish Language in the Digital Age

2.1 General Facts

Irish is the first official and national language of the Republic of Ireland, with English as the second official language. Irish Sign Language has official legal recognition since 2017.[3] Figures from the 2016 census report that 39.8% (1.7 million) of the population can speak Irish, while only 1.5% (73,000) speak Irish on a daily basis outside the education system. Irish is also recognised as a minority language in Northern Ireland and has been an official language of the European Union since 2007. The lifting of a derogation on official EU translations at the end of 2021 led to Irish becoming a full working language of the EU. Despite this status, according to the UNESCO Atlas of the World's Languages in Danger, Irish is considered "definitely endangered" (Moseley, 2012).

Irish has three main dialects and a number of sub dialects. The major dialect divisions follow the demarcation of the three provinces of Connaught, Munster and Ulster. These dialects differ at many levels in terms of their sound system, prosody, vocabulary and structural features. The written form of the Irish language was standardised in 1958 with the publication of *An Caighdeán Oifigiúil* 'The Official Standard',[4] which draws on the individual dialects to provide the standardised spelling and grammar to be taught in schools. However, there is no spoken standard variety. The native dialects are equally deemed standard, which has implications for speech technology development. Latin script is used for the language's writing system, with an alphabet similar to English, but excluding j, k, q, v, w, x, y, z (except in loanwords). However, the consonants are not marked for the fundamental contrast of palatalisation and velarisation of Irish. All long vowels are accented as follows: á é í ó ú.

Linguistically, Irish shares distinctive features with other Celtic languages such as a verb-subject-object (VSO) word-order and rich morphology (Stenson, 1981). Adjectives and other modifiers usually follow noun phrases, both a copula and substantive verb 'to be' are used, while clefting of nominal, adverbial and prepositional phrases is frequent, influencing the design of chunking and parsing tools (Uí Dhonnchadha, 2009; Lynn, 2016). When paired with English, their divergent word orders can pose challenges for applications such as alignment tools and machine translation (MT) systems (Dowling et al., 2019). Inflection in Irish mainly occurs through suffixation, but initial mutation through lenition and eclipsis is also common (The Christian Brothers, 1988). Verbs are inflected for tense, number and person, while nouns are inflected for number and case. Nouns are either masculine or feminine in grammatical gender. As with other Celtic languages, prepositions can inflect for person and number. This overall inflectional nature leads to sparsity in Irish datasets, which has

[3] Irish Sign Language (ISL) is not based on the Irish language however, and is reported to be closely related to French Sign Language with influences from British Sign Language. 'Spelled-out' words are based on English.

[4] The most recent version, published in 2017, is available at https://data.oireachtas.ie/ie/oireachtas/caighdeanOifigiul/2017/2017-08-03_an-caighdean-oifigiuil-2017_en.pdf

been seen to impact Natural Language Processing (NLP) development within an already low-resourced context (Lynn et al., 2013).

In terms of usage, there are dispersed 'Gaeltacht' regions across Ireland where Irish is spoken daily as a first language in the community (parts of counties Cork, Donegal, Galway, Kerry, Mayo, Meath and Waterford). These communities are scattered geographically and as such, their dispersed nature dilutes the density of Irish language speakers across the country. In addition, English is becoming increasingly used in these Gaeltacht regions, partially due to its monopolising digital presence, which is impacting the speaking habits of those who are digitally-connected (Ó Giollagáin and Martin, 2015). Outside Gaeltacht regions, Irish is also spoken at home by many families in urban areas, where many also try to increase their proficiency in the Irish language to make it a second working language in their day-to-day lives. This is reflected through various initiatives that have been set up to facilitate connecting Irish speakers, such as Borradh[5] (aimed at building professional networks through Irish) and the Pop Up Gaeltacht.[6] Both use social media to connect members, and have witnessed a rise in the number of urban-based Irish language speakers finding new ways to connect and use Irish in social settings. Finally, with Ireland's long history of emigration, it is unsurprising that there is a large diaspora of Irish speakers across the world, with particular concentration in countries such as the UK, US, Canada and Australia (Connolly, 2021). Over 40 universities and other third-level colleges worldwide are benefiting from the Irish Government's financial support in developing the teaching of Irish abroad through programmes such as the Fulbright Commission in Ireland.[7]

Both the education system and government support and policies play significant roles in the continued transmission of the language in Ireland. In terms of education, Irish is a compulsory core subject in the school curriculum at primary and secondary level, ensuring that the general population has considerable exposure to the language. Research has found, however, that there has been a decline in Irish language proficiency of students of English-medium schools (Harris, 2006). On the positive side, outside of the Gaeltacht regions, the number of Irish-medium pre-schools, primary and secondary schools is growing across both the Republic of Ireland and Northern Ireland (Ó Duibhir et al., 2017). Language learning online has grown more popular in recent years. Platforms such as Fáilte ar Líne[8] at Dublin City University (DCU) have proven successful in this respect, along with the Duolingo language learning app.[9] However, there is a severe lack of sophisticated Computer Assisted Language Learning tools (CALL systems) for Irish, relegating Irish learning activities in classrooms to outdated methods.[10] With respect to education for Irish Language Technology, there has been a notable lack of training and education programmes that combine skill sets of both the Irish language and technology. This has been particularly evident at third level, where there is only one LT-focused undergraduate course available in Ireland.[11] This programme offers a choice of language options, but the number of students taking the Irish language option is very low.

In terms of government policy, The Official Languages Act (2003) has the primary objective of ensuring the improved provision of public services through the Irish language. The Office of *An Coimisinéir Teanga* (Language Commissioner) was established under the Act in 2004 to monitor compliance by public bodies with the provisions of the Act, and to take ap-

[5] https://www.borradh.ie
[6] https://en.wikipedia.org/wiki/Pop-Up_Gaeltacht
[7] https://www.gov.ie/en/publication/08d6f-third-level-education-overseas/
[8] https://www.failteonline.ie
[9] As of December 2021, figures provided by Duolingo indicate over 1.15 million active learners on the Irish language app: https://www.duolingo.com/course/ga/en/Learn-Irish
[10] It should be noted that the uptake and response to the recently developed bespoke Irish iCALL platform, An Scéalaí, is demonstrating an acute appetite for Irish-specific CALL platforms (Ní Chiaráin and Chasaide, 2019).
[11] B.A. in Computer Science, Linguistics and a Language (CSLL) https://www.scss.tcd.ie/undergraduate/computer-science-language/

propriate measures to ensure such compliance. In addition, the 20 Year Strategy[12] for the Irish Language 2010-2030 recognises the State's commitment to the language's revival, while The Action Plan for the Irish Language (2018-2022)[13] provides a framework that focuses on specific and realistic actions to be implemented in the given time frame.

Mainstream media also plays a strong role in strengthening and supporting the lives of those who choose to use Irish in their daily lives. The national Irish language public television network, TG4, provides both a dedicated Irish language content television channel as well as extensive online media, while the state broadcaster, Raidió Teilifís Éireann (RTÉ), has a statutory duty to broadcast and publish some Irish content, including daily news bulletins. Likewise, Irish language radio stations, Raidió na Gaeltachta (national) and Raidió na Life (Dublin area), provide Irish language audio content through live broadcasting, website content and podcasts.

2.2 Irish in the Digital Sphere

According to the Central Statistics Office's 2019 figures, 91% of the population have internet access.[14] Based on figures from 2020, it is estimated that there are around 300,000 websites with the Ireland-based `.ie` domain and that Irish language content is found across roughly 1,500 (0.5%) of these domains.[15] Thanks to government policies, there is a growing amount of online content found across the websites of public bodies, universities, language schools, and language organisations, etc., as observed through web-crawling data collection efforts in Ireland under the European Language Resource Coordination (ELRC), the Paracrawl project (Bañón et al., 2020) and through the Crúbadán project (Scannell, 2007). However, there are still low numbers of businesses localising their websites to an Irish-speaking market. This lack of engagement through Irish in business is also reflected in the minimal presence of Irish language content on professional networking sites such as LinkedIn. On the other hand, the Irish Wikipedia (An Vicipéid) is a growing and valuable resource of Irish digital content. For the past several years, it has ranked between 90th-93rd in terms of size and number of articles.

The use of Irish in social media has grown steadily over the past several years and is now prevalent across platforms such as Facebook, Twitter, Instagram, etc. (Lackaff and Moner, 2016). There is a strong community of active Twitter users, with over 4 million Irish tweets posted to date.[16] Both Facebook and Twitter are often used by both groups and individuals to disseminate information relating to the Irish language such as events, government policies, education, language learning and so on, with the hashtag #Gaeilge often used to tag and categorise such posts (Nic Giolla Mhichíl et al., 2018). However, in contrast to the large presence of Irish language users on these platforms, there is still minimal support for Irish from technology companies. The quality of Irish translations from Google Translate and Bing Microsoft Translator still prove unreliable for many within particular domain settings, and much controversy has arisen around the frequent misuse of automated translation processes that bypass human post-editing or verification.

Open-source software such as Firefox, Thunderbird, GNULinux, LibreOffice and KDE have all been localised into Irish by volunteer translators.[17] Also thanks to volunteer translation efforts, Google offers a localised version of the interfaces for Gmail and Google Search, but the other components of Google Workspace and YouTube are not localised. Crowdsourcing

[12] https://www.gov.ie/en/policy-information/2ea63-20-year-strategy-for-the-irish-language/
[13] https://www.gov.ie/en/policy-information/1418a-action-plan-2018-2022/
[14] https://www.cso.ie/en/releasesandpublications/ep/p-isshh/informationsocietystatistics-households2019/householdinternetconnectivity/
[15] Information provided by Prof Kevin Scannell, St Louis University, based on web-crawling activities.
[16] http://indigenoustweets.com Figures as of January 2022
[17] https://riomhacadamh.wordpress.com

translation efforts have led to an unverified localisation of Facebook, which also does not yet offer the option to translate Irish language posts. Until recently, Twitter offered tweet translations through Bing. Yet, since switching to Google Translate as a translation partner in 2019, tweets are no longer identified as being written in Irish and this translation option has disappeared.

3 What is Language Technology?

Natural language[18] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

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

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

With its starting point in the 1950s with Turing´s renowned intelligent machine (Turing, 1950) and Chomsky´s generative grammar (Chomsky, 1957), LT enjoyed its first boost in the 1990s. This period was signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in machine learning, rule-based systems have been 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

[18] 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).

computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches.

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

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

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

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[19]

[19] 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://tinyurl.com/2p9ed6tp). 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).

4 Language Technology for Irish

Overall, support for the development of Irish Language Technologies has not changed significantly over the past 10 years. That is to say, that there is still a large number of fundamental tools and datasets required to build these technologies that not yet available for Irish (see Table 1 in Section 5). Some progress has been made, however, in the area of text analytics and MT, thanks to data collection and corpus creation through a number of short term projects based in academic institutions, funded by EU-projects and national funds, or self-funded. Steady progress has been made in terms of developing speech resources and technology, particularly in speech synthesis for the three main dialects, while further dialects still need to be catered for. Applications have been developed to make these voices available to the public, e.g. in accessibility aids such as a screenreader for the visually impaired, and they have also been integrated into computer assisted language learning (CALL). While steps have been made towards speech corpus development, there is still no market-ready automatic speech recognition system available for Irish – either through proprietary software or open-source software. Fundamental building blocks such as syntactic analysis tools have progressed, but in terms of technological readiness, the underlying datasets are still too small to build robust application-ready systems. From a natural language understanding perspective, there is a severe lack of semantic-based datasets and tools.

It should be noted that regular version updates of some multilingual datasets recorded on the ELG can result in inflating overall figures for Irish (e. g. Universal Dependencies, ParaCrawl), as well as multilingual datasets of which only a small proportion represents Irish (e. g. Multilingual corpus from the Publications Office of the EU on the medical domain). The real picture, however, is that Irish is very much a low-resourced language. The following summarises the type of language resources and tools that currently exist for Irish.

4.1 Language Data

Monolingual Corpora

There is a limited amount of freely available monolingual corpora for Irish. Historically, much of the corpus development for Irish in Ireland was geared towards the purposes of dictionary development, such as The New Corpus for Ireland–Irish (NCII) (Kilgarriff et al., 2006), which consists of over 30 million words of widely mixed domain (e. g. newswire, literature, legal). It has been automatically part-of-speech (POS)-tagged and has proven valuable in the development of dictionaries, MT systems, language modelling, to name but a few. However, due to the copyright nature of most of the sources that make up this corpus, it is only available for research purposes under restrictive licensing. Furthermore, some of the content does not reflect contemporary language usage or style (e. g. literature and prose). As such, its use as training data for developing modern LT systems is rather limited. On the other hand, the Gaois Corpus of Contemporary Irish (Ní Loingsigh et al., 2017) contains up-to-date content from news media and e-zines, yet it is also restricted in terms of access and usability due to copyright.

Most other Irish monolingual text corpora have been products of specific research projects or PhD theses. For example, the first gold-standard POS-tagged corpus (Uí Dhonnchadha, 2009), the Irish treebanks (Lynn, 2016), a corpus of idioms (Ní Loingsigh, 2016), the EduGA Corpus of Educational Materials (Meachair, 2020), and a multi-word expression (MWE) corpus (Walsh et al., 2020) were all outputs of PhD research. Most of these corpora are still relatively small and are not yet large enough to train high accuracy models.

Very few domain-specific monolingual datasets exist. For example, both the EduGA and TEG Learner Corpus of Irish are language education domain corpora. The TEG corpus, con-

taining text from learner proficiency tests and including POS-tags, contains manual error-correction markup, with potential for use in adaptive learning systems. A learner corpus is also being created, An Corpas Cliste, based on input from users of the iCALL learning platform, An Scéalaí (Ní Chiaráin and Ní Chasaide, 2020). In terms of varying genre, the Irish UD Twitter treebank (`TwittIrish` (Cassidy et al., 2022)) is the only user-generated content corpus. It is both tagged at a morpho-syntactic level and contains code-switching information. The Comhrá Spoken Corpus (Uí Dhonnchadha et al., 2012) contains about 240k words of transcribed spoken language from all of the major dialects. Transcriptions are segmented, POS-tagged and aligned with audio files.

The Irish Wikipedia (*An Vicipéid*) has seen consistent growth over the past several years. With the recent appointment of a project coordinator for Wikimedia Ireland, some wiki-editing training workshops for Irish speakers have taken place. The dataset has proven useful in the development of resources such as Multilingual BERT (Devlin et al., 2019) and the Irish gaBERT language model (Barry et al., 2021).

Apart from some scientific content on *An Vicipéid*, it should be noted that there is a considerable lack of Irish monolingual corpora available for specific domains (e. g. legal, medical, etc.). In terms of the types of linguistic annotation, only tagging at the morpho-syntactic level is available. That is, there are no suitably tagged corpora annotated for anything other than basic processing, for example there is nothing that could be used for semantic parsing, question-answering, sentiment analysis, named entity recognition, entity extraction, anonymisation or discourse analysis.

Bilingual/Parallel data

Significant advancements have been made in the collection and availability of bilingual texts for the purposes of English<>Irish machine translation, largely due to Ireland's involvement in the European Language Resource Coordination (ELRC) project. Much needed awareness-raising and education around the usefulness of translation technology, as well as the importance of effective translation data management in the public sector was made possible through national outreach activities such as ELRC workshops and onsite visits to public administration bodies. In each case, the backing and support of the EU and *Roinn na Gaeltachta* (DTCAGSM)[20] provided the necessary weighting behind the national data collection campaign. As much of the data collected through the ELRC efforts in Ireland were sourced from public bodies, the majority of this data collection is available to the wider public under the EU Open Data Directive. Domains covered include general public administration, eJustice and eProcurement.

Irish parallel text is amongst a number of multilingual resources created at the EU level (e. g. DGT translation memory, COVID-19 EUR-LEX, etc.) released through ELRC-SHARE,[21] as well as multilingual resources created through wide-scale crowdsourcing (OSCAR) or web-crawling (ParaCrawl).[22] Medical domain Irish data collected thus far is limited to multilingual datasets such as COVID-19 and the EU Vaccination portal data, along with a monolingual corpus from the health.gov.ie web site.

It is also worth noting that Irish is one of a large collection of comparable corpora (both spoken and written) under development as part of the ICC International Comparable Corpus project (Čermáková et al., 2021).

[20] The Irish government department (herein referred to as *An Roinn*) responsible for Irish language affairs – currently housed by The Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media and subject to change https://www.gov.ie/en/organisation/department-of-tourism-culture-arts-gaeltacht-sport-and-media/

[21] https://elrc-share.eu

[22] It should be noted that the figures for Irish web-crawled corpora noted in the ELG catalogue may be inflated due to the number of overlapping projects (such as DCU Irish MT data on ELRC-SHARE, ParaCrawl, Crúbadán).

Multimodal Corpora

Different kinds of speech corpora have been collected in connection with the ABAIR initiative, including the ABAIR General Speech Synthesis Corpus (Kelly et al., 2009). Recordings of native speakers from the three main dialects have provided the foundation of the ABAIR TTS systems (Ní Chasaide et al., 2017), and these are being extended to include further subdialects. There are currently around 25 hours of recorded speech available, of which selected portions have been processed for synthesis – edited, segmented, and aligned to the phonetic transcription (X-SAMPA, IPA) with stress marking. In complementary work, ongoing development of the ABAIR Compact Speech Synthesis Corpus (An Corpas Beag) thus far involves (around 2K) scripted prompts for maximal phonetic coverage with minimal material, and contains the same level of annotation.

Further corpora involving both live recordings and crowdsourced recordings of predominantly native speakers using the online facility Míle Glór[23] are being collected (25 hours to date) and processed for the development of an automatic speech recognition (ASR) system. In addition, the Mozilla Common Voice project has so far collected 420 minutes of Irish speech (both native and non-native speakers) through crowdsourcing efforts.

There are currently only two multimodal datasets containing text and images. The CaptionCommons Corpora contain bilingual captions in English and Irish, collected from 434 images in Wikimedia Commons (compared to 3,560 for Basque). In addition, the Massively Multilingual Image Dataset (MMID) contains paired images and words for over 90 languages, including Irish, collected through Mechanical Turk (Pavlick et al., 2014). While the quality of the latter dataset has not yet been verified for Irish as the translations were crowd-sourced and potentially machine-translated, it could be used as a starting point for the development of an image captioning tool.

Lexical Resources

For a minority language, Irish is relatively well-resourced when it comes to linguistic knowledge bases such as electronic dictionaries, terminology databases, thesauri, gazetteers and glossaries. Most dictionary development (monolingual and bilingual) has been funded by Foras na Gaeilge.[24] Some are accessible from a single domain[25] or through phone apps. Due to copyright restrictions, however, most of them only offer single user queries or data access for research purposes only. The National Morphology Database (Měchura, 2014) is hosted at the same site and is a large (currently 43,000 entries) and valuable open-source collection of Irish words with information on their inflected forms and various other linguistic properties.

Another freely accessible resource is the National Terminology Database developed by Fiontar & Scoil na Gaeilge (DCU).[26] This large database (currently 185,000 entries) is referenced by the general public, students, freelance translators and translators at EU institutions alike. The success of this database is based on the existence of a Terminology Committee for Irish (An Coiste Téarmaíochta)[27] who develop, approve and provide authoritative standard Irish terminology. Equally popular is the Pota Focal[28] site which hosts a number of resources such as a dictionary, a glossary, as well as a verb valency dictionary and thesaurus, the latter of which is powered by *Líonra Séimeantach na Gaeilge* (LSG) – an Irish Wordnet.[29]

[23] https://phoneticsrv3.lcs.tcd.ie/studio/ga/recorder/
[24] The public body primarily responsible for the promotion of the Irish language across the island of Ireland.
[25] https://www.teanglann.ie/ga/
[26] https://www.tearma.ie
[27] https://www.tearma.ie/eolas/coiste.en
[28] http://www.potafocal.com
[29] https://github.com/kscanne/wordnet-gaeilge

Models and Grammars

A constraint grammar for Irish is available, which forms the basis of a chunking tool (Uí Dhonnchadha, 2009). In addition, the rule-based computational grammar library for Irish – Gramadán – accompanies the Irish National Morphology Database to allow for the generation of inflected forms of nouns, adjectives, verbs and prepositions.[30]

Thanks to the existence of open-source raw corpora such as Wikipedia, Irish has been included in the latest suite of Transformer Language Models (BERT, M-BERT, BERT sentence Encoder – LaBSE), and through additional research at DCU using the NCII, Wikipedia, web-crawled data and ELRC-related data collections, there is now an Irish gaBERT language model available (Barry et al., 2021). Word2Vec and ELMO embeddings have also been made available through the NLPL[31] based on the Irish UD treebank.

4.2 Language Technologies and Tools

Irish is lacking in general in the availability of robust speech and language tools. The ELG catalogue lists a number of multilingual tools and services that include Irish as a supported language (e. g. Bitextor, Opus MT, Systran) and some others that offer a service (e. g. LIMA, NLP-Cube, GNU Aspell) based on the same underlying datasets or tools (e. g. UD_Irish-IDT treebank, Gaelspell). In the summary below, the focus is on tools specifically developed and designed for Irish.

Text Analysis

The most significant development in Text Analysis tools for Irish are the XFST Finite State suite of tools that include a tokeniser, lemmatiser, morphological analyser and POS-tagger (Uí Dhonnchadha, 2009). This collection of fundamental tools have been the building blocks for a number of other tools and resources. There are also dependency parsing models available through UDPipe and Stanza, for example, that are based on the two Irish Universal Dependency (UD) treebanks – general domain and Twitter content. However, their accuracy is not yet reliable for downstream tasks until the treebanks grow in size.[32] Neither have been evaluated in downstream applications. Parsing Irish tweets is particularly challenging given the small treebank available and the propensity for noise in user-generated content. Early stage research is underway at DCU in the development of a tool that will also process Irish multiword expressions.

There is only one open-source spell-checker (GaelSpell) and one open-source grammar checker (An Gramadóir).[33] These provide the underlying technology for a number of free proofing services provided online. Very few proprietary tools or applications offer predictive text or auto-correct for Irish. An Irish-supported language identifier is also available from the Crúbadán project (Scannell, 2007).

Speech Processing

Text-to-speech systems for the three main Irish dialects have been available for some time through the ABAIR initiative (TCD) and work is ongoing to provide coverage of the further dialects. These systems are proving invaluable in many domains of application – for the

[30] https://github.com/michmech/Gramadan
[31] http://vectors.nlpl.eu/repository/
[32] The size of the Irish_UD-IDT currently stands at 4910 trees, and when used to train a neural parser (using gaBERT), achieves an accuracy of 85% Labelled Attachment Score.
[33] https://cadhan.com/gaelspell/sios-en.html

general public, for language learners, and for those with disabilities. For the general public, the ABAIR website[34] provides access to the synthesis systems, including a web reader which allows any digital text to be read aloud. Language learning games and platforms are being developed, e.g. the An Scéalaí iCALL platform.[35] For those with visual impairment, a screen-reading facility with both spoken Irish and Braille outputs is available.[36] Many more applications are needed, such as assistive technologies for the non-verbal, and some initial work is underway here.

While an Irish Automatic Speech Recognition ASR system is under development and a fledgling beta version is available online, ÉIST,[37] there is still no reliable ASR system available, either through proprietary (e. g. Alexa) or open-source software.

Translation Technologies

Thanks to data collection and both statistical and neural MT research carried out at DCU (e. g. through PhD research (Dowling et al., 2015, 2020; Lankford et al., 2021)), SMT and NMT systems have been developed for English-Irish. Through the Connecting Europe Facility (CEF) PRINCIPLE Project,[38] bespoke MT systems were developed for a number of public bodies such as the Department of the Gaeltacht, Rannóg an Aistriucháin, Foras na Gaeilge and the National University of Ireland Galway (NUI Galway). These systems were designed to support translators in a professional translation workflow. In addition, a self-funded open-source Irish-Scottish Gaelic SMT system has been developed (Scannell, 2014).

As an official EU language, Irish is included amongst the languages supported by eTranslation, the European Commission's MT platform. Through collaboration between DCU and the EU's Directorate-General for Translation (DGT), along with the ELRC data collection efforts, the quality of Translation for Irish has been increasing steadily over the past few years. This is of significant relevance in the context of Irish becoming a full working EU language in January 2022. Google, Bing, and the IRIS MT system (Arcan et al., 2016), developed at NUI Galway, all support Irish as free general purpose online translation services.

Gaps

Mainly due to the lack of underlying data resources, dedicated funding and skill-sets, to date there has been little system development for Automatic Speech Recognition, and no system development for Automatic Subtitling, Information Retrieval, Information Extraction, Natural Language Generation, Semantic Role Labelling, Named Entity Recognition, Sentiment Analysis, Question-Answering, Virtual Agents, Adaptive Learning, Entity Extraction and Linking (knowledge bases) or Personal Identifiable Information (PII) detection.[39]

4.3 Projects, Initiatives, Stakeholders

When discussing the language technology landscape in Ireland it is important to differentiate between general LT-developments, which are focused on English, and specific projects and initiatives dedicated to supporting the Irish language. For example, the National AI Strategy for Ireland,[40] entitled "AI – Here for Good", sets out to provide a high-level direction to the

[34] https://www.abair.ie
[35] https://www.abair.ie/scealai
[36] https://abair.ie/en/accessibility/
[37] https://phoneticsrv3.lcs.tcd.ie/rec/irish_asr
[38] https://principleproject.eu
[39] While still at early stages and of low accuracy, some steps have been taken towards PII detection in the CEF-funded MAPA Project https://mapa-project.eu.
[40] https://www.gov.ie/en/publication/91f74-national-ai-strategy/ See page 42.

design, development and adoption of AI in Ireland. The strategy makes some reference to LT in general. However, the focus for LT relates to English language based AI, while the following minimal acknowledgement is made to the need for LT support for Irish (in the public sector only): "To render AI systems accessible to a wider range of our population, as well as to develop services in Irish based on AI for Irish language-speakers, good language technology resources need to be developed".

While there are extensive LT industry bodies and research centres in Ireland (e. g. Apple, Accenture, Google, SoapBox Labs, AYLIEN, CeADAR, ADAPT Centre) their primary focus so far, with respect to Irish customers or members of the public, is on supporting the English speaking population only. Aside from some localisation firms supporting Irish translations, there are no Irish-language specific industry players in the speech and language technology space. Irish-language related projects are therefore mostly managed and supported through *An Roinn*'s Irish Language Support Schemes[41] and Foras na Gaeilge.

In a positive step forward, *An Roinn* is overseeing the development of a vital Digital Language Plan that will outline the need for R&D in speech and language technologies for Irish. This document is under review and awaiting publication. In addition, with a new round of Science Foundation Ireland Funding, the ADAPT Centre (phase II) research strategy plans to expand in order to address the issue of low-resource languages, which hopefully will include Irish LT.

Many of the achievements and advancements in the development of language data and tools for Irish have been as a result of small, short term funded and self-funded projects or PhD theses (e. g. spell-checker, grammar-checker, morphological database, POS-tagger, treebanks, etc) and more recently EU funding related to MT data collection. Despite the lack of an R&D roadmap, dedicated funding or established infrastructure for Irish LT, there have been a number of notable projects that have helped shape the LT landscape into what it is today.

The longest running LT project for Irish is the ABAIR initiative, at Trinity College Dublin (TCD), which has provided speech synthesis for the 3 main dialects and and is increasingly providing downstream applications such a screen-reader and CALL systems. In terms of Automatic Speech Recognition (ASR), development is underway of the Míle Glór ('A Thousand Voices') Speech Recognition Corpus. The initiative aims to collect at a minimum 1,000 voices of Irish speakers of all ages and dialects.

The CEF-funded projects have proven to be most impactful in the area of machine translation. The European Language Resource Coordination (ELRC) has provided the much needed backing for this previously untapped data collection and opportunity for awareness-raising through outreach workshops (2016, 2017, 2021).[42] Of significant importance was the establishment of Ireland's National Relay Station (NRS)[43] through the CEF-funded European Language Resource Infrastructure (ELRI)[44] project, which has seen many representatives from public organisations upload their own legacy and current translation data to the portal, from where the data is shared with other users and relayed onto ELRC-SHARE for the purpose of improving the English-Irish MT engine of the eTranslation system. A long term sustainable funding model is required to ensure continued collection of up-to-date translations from across the public sector both in the Republic of Ireland and Northern Ireland.[45]

Irish has been identified as a low-resource language in terms of the amount of parallel data available, and in terms of quality and usability of the Irish eTranslation system. Partnered with similarly low-resource languages (Croatian, Icelandic and Norwegian) the CEF-funded PRINCIPLE project saw the collection of data for the eHealth and eJustice domains, while

[41] https://www.gov.ie/en/publication/7547d-language-support-schemes/
[42] https://lr-coordination.eu/pworkshops
[43] https://elri.dcu.ie/ga-ie/
[44] http://www.elri-project.eu
[45] The portal is managed by researchers at DCU, with a further 2 years' funding provided by *An Roinn* until 2023.

providing brief access to bespoke MT systems for early adopter stakeholders to demonstrate the benefits of using Irish MT in the public sector.

In terms of Text Analysis, the GaelTech project (2017-2023) at DCU involves the training of PhD postgraduate students and linguists in Irish language NLP. The specific focus of this project is POS-tagging, syntactic parsing, language modelling and the processing of user-generated content, code-switching and multiword expressions. All datasets (annotated corpora, lexicons, models) and tools are being made available under open-source licensing. The project is also funded by *An Roinn*.

In a step to address the shortcomings of the NCI and its accessibility issues, the development of the National Corpus of Ireland[46] is underway as of January 2022 by researchers at Fiontar & Scoil na Gaeilge in DCU and TCD. This large national corpus of contemporary Irish, encompassing both written and spoken sources, will be made accessible under open-source licensing to both the research community and members of the public. The written Irish portion of this raw corpus is set to be around 100 million words. Built specifically with language technology in mind, the intention is to publish resources such as word-frequency and ngram lists, as well as language models. The project is also funded by *An Roinn*.

Irish LT researchers are also part of the Celtic Language Technologies research group, which organises academic workshops (CLTW)[47] allied to major international conferences in the LT field, publishing papers in relevant peer-reviewed proceedings.

5 Cross-Language Comparison

The LT field[48] 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 previously unforeseeable. 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[49] 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)

[46] https://www.corpas.ie/en
[47] See for instance https://aclanthology.org/volumes/W14-46/
[48] This section has been provided by the editors.
[49] 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.

  - 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 ≥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[50]

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[51] 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

[50] 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.

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

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.[52]

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

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

5.4 Results and Findings

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

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g. German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[53] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

While a fifth level, *excellent support*, could have been foreseen in addition to the four levels described in Section 5.2, we decided not to consider this level for the grouping of languages. Currently no natural language is optimally supported by technology, i. e. the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning,

[52] 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.

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

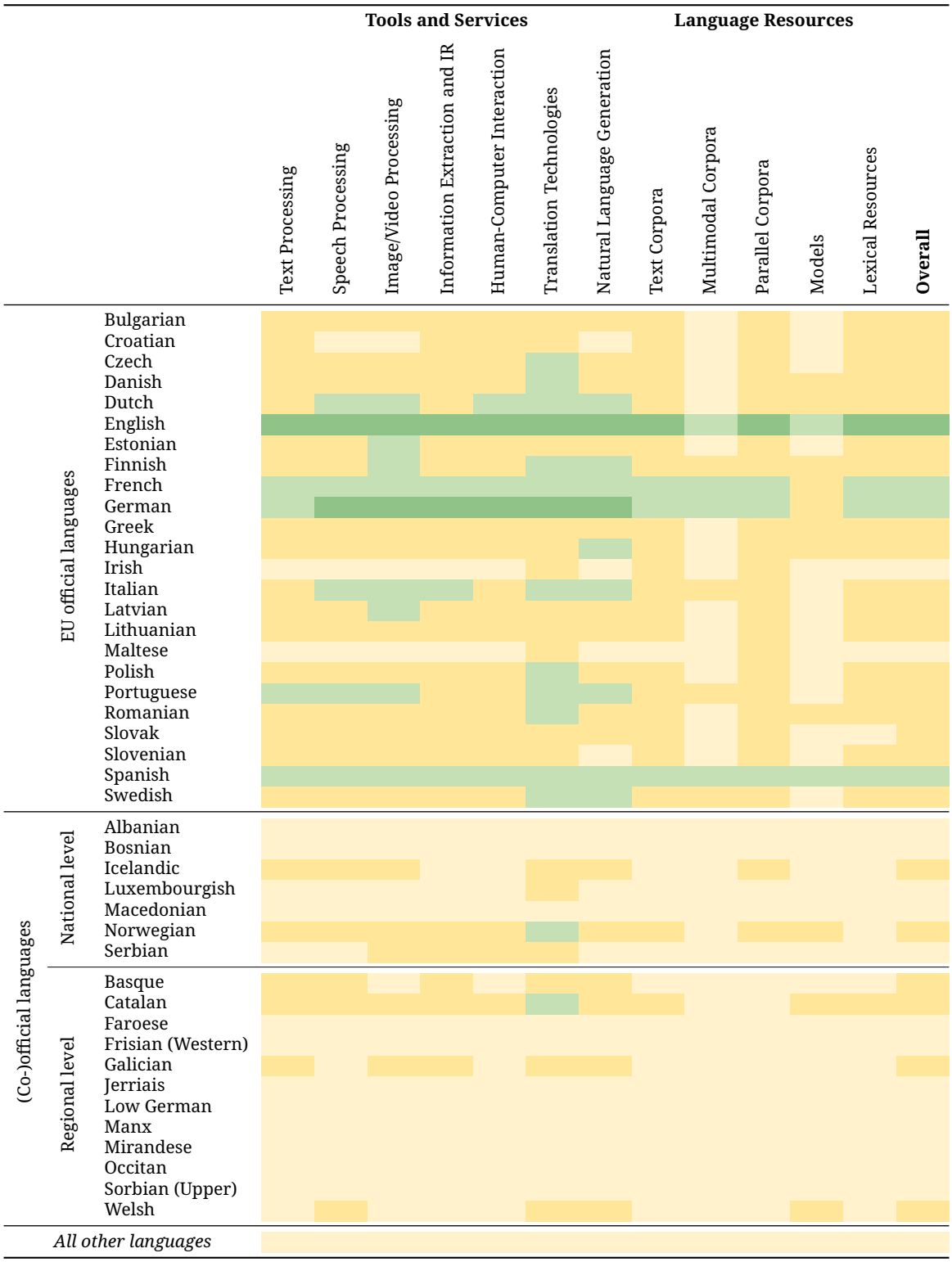

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

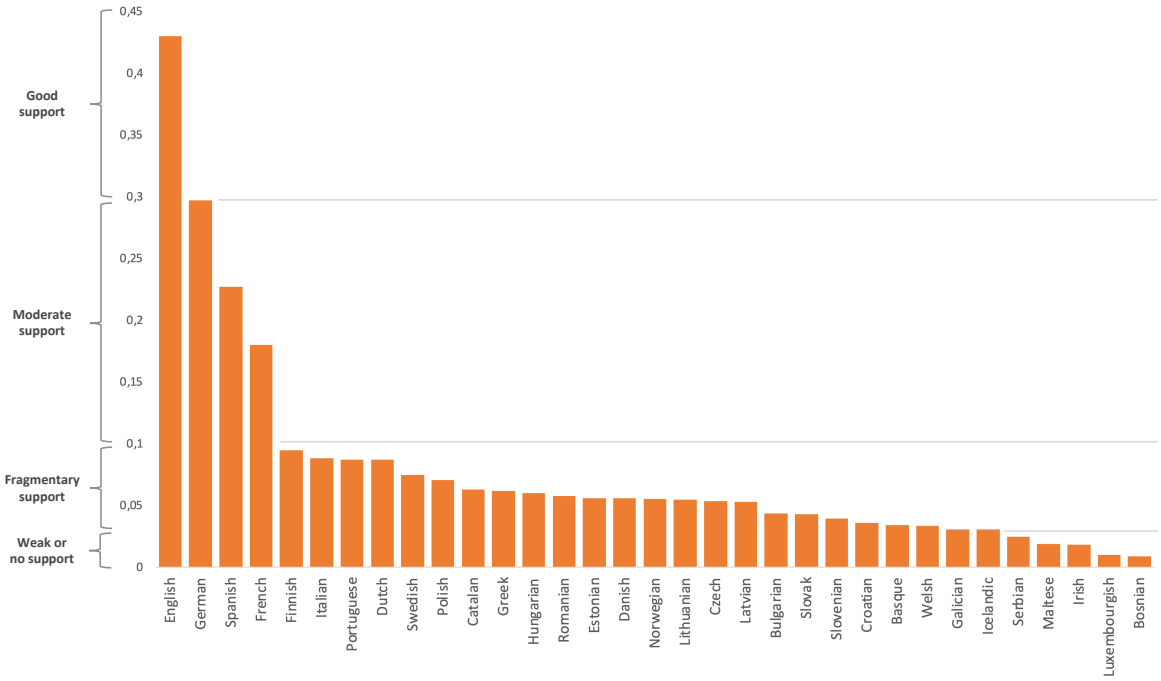

Figure 1: Overall state of technology support for selected European languages (2022)

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

In Ireland, Artificial Intelligence is already a part of everyday life. While unknown to many, language technology plays a key part in this. We use it when typing messages, emails or posts, browsing the internet, shopping online, filtering emails, posting reviews, communicating,

interacting with smart devices and appliances (internet of things), etc. This technology is only set to continue and evolve. Yet it is important to take stock of the fact that, for Irish citizens, most of these technologies can only be accessed through English. For those speaking Irish as a first language, a shift to English is currently commonplace in order to stay digitally connected. This language shift can occur based on the lack of the most basic technologies, such as proofing tools (e. g. autocorrect changing an Irish word to English due to lack of support). As these advances continue and unless significant action is taken to ensure digital equality, Irish will be left further behind and become less frequently used.

While there have been a number of developments in speech and language technology over the past 10 years, many commonly used and necessary technologies are still not available (e. g. ASR (Automatic Speech Recognition), Dialogue Systems, Information Extraction, Named Entity Recognition, Automatic Subtitling). Similarly, while there has been an increase in focus on the creation of open-source corpora, not all have been specifically designed for the purpose of developing data-driven LT systems. As such, much of the corpora developed or in development are raw and unannotated, yet with much potential should further investment be made into developing relevant tools.

Some observations can be made with respect to future strategies to tackle these shortcomings.

Change of focus

To date, the Irish language has received much investment into the development of dictionaries and terminologies. This has mainly been driven by a focus on supporting translators and Irish language learning. However, a shift in focus is required to recognise technology as an equally important axis for continued language use. This shift should see a broadening of scope in terms of funding within the wider lens of speech and language technology. Some areas may hold priorities over others in terms of urgency (e. g. CALL, speech recognition, text analysis may take priority over chatbots). Such funding opportunities should be made available to Irish language organisations, education institutions, research centres and entrepreneurial groups in both the Gaeltacht and non-Gaeltacht communities.

Untapped Potential

The value of language data is broadly unknown amongst Irish citizens and across the Irish public sector (Berzins et al., 2019). There is much untapped yet currently inaccessible data that could make a huge impact on the future of Irish LT if collected and applied appropriately. For example, there is much aligned audio and subtitle text data available in the archives of the national broadcaster (RTÉ) and Irish language broadcaster (TG4) that could easily be used to build ASR and automatic subtitling systems. The national placenames and biographies databases[54] and the Gaois Database of Irish-language Surnames[55] could be leveraged to build a named entity recogniser. The available language learning corpora could be used as a basis for developing efficient CALL systems. In addition, the creation of Irish content online is a mix of both curated and user-generated content. YouTube has proven to be a popular platform for promoting Irish language through the arts and education, with content in the form of edutainment, influencer channels and vlogs. This data could be collected and processed to develop user-generated corpora necessary to build tools that could process modern written and spoken Irish. Furthermore, the general positive disposition and altruistic nature of Irish speakers toward supporting the language lends themselves to the leveraging of citizen science or crowd-sourcing approaches to data collection, dataset creation

[54] https://www.logainm.ie/en and https://www.ainm.ie
[55] https://www.gaois.ie/en/surnames/info/

and tool evaluation. This has already proven successful in the Meitheal Dúchas manuscript transcription project.[56]

Need for Dedicated LT Programmes

Major challenges faced by many past Irish LT projects were related to skill shortages. Due to the lack of dedicated education and training programmes in this field, it has proven difficult to source researchers, linguists or engineers with the right combination of skills (e. g. Irish language, computer science, linguistics). The untapped potential of the resources described in the previous paragraph highlights clearly how much more progress can be made if the right skill sets can be found. In the short-term, further support is required for the Irish stream of the B.A in CSLL course offered at TCD (see Section 2). In addition, scholarships could be offered to those taking up general AI or Data Science postgraduate courses, with the incorporation of an Irish LT practicum. The inter-disciplinary skills gap may also gradually be addressed through recent initiatives such as Clár Techspace,[57] which offers training opportunities in technology literacy (both through English and Irish) through local youth organisations. Thanks to such initiatives, along with the introduction of Computer Science as a subject at Leaving Certificate level, is hoped that the pool of available skills will widen.

Long-term strategy

In the absence of a Digital Language Strategy, as yet, there are no long term funding schemes or research centres dedicated to Irish LT. A change in this regard will ensure: a strategic plan for safeguarding Irish in a digital age, support for dedicated LT education and training, investments in data collection and annotation, development of sophisticated LT tools and services that are production ready and easily integrated into smart devices or online applications.

Open-source culture

Open-source describes data or source code that is freely available to use, modify, and redistribute. In terms of technology, this usually means that the code or tool is designed in such a way that it can be easily integrated into other systems without the need for specific licensing or ongoing support. As noted above, there are many high quality resources available for Irish that are under strict copyright protection, rendering them unusable for general purpose. However, the power of open-source is reflected in the fact that the Irish gaBERT model has been downloaded at a rate of 500 downloads per month.[58] It can be easily concluded that most of these downloads have been made by researchers outside of Ireland, which demonstrates the power of data and tool sharing. It is important therefore, where possible, that all data and tools developed for Irish are made available under open-source licensing. This will ensure that their use is not restricted to small institutions that might not have the skills or resources to further develop them for application use.

Collaboration

We have already seen the benefits of collaborations in the advances of LT for Irish. This applies both to academic collaborations and cross-government collaborations. In particular, we can see how EU-funded projects have allowed for sharing of resources, knowledge and

[56] https://www.duchas.ie/en/meitheal/
[57] https://kinia.ie/clar-techspace/
[58] As per January 2022. https://huggingface.co/DCU-NLP/bert-base-irish-cased-v1

expertise in order to reach data collection or system development goals. In academia, sharing knowledge and experience with international institutions have been integral to much research and development. Collaborations on a national level are also important, for example summer internships could be offered to Clár Techspace pupils at universities, graduate internships could be made available at large Ireland-based technology companies or international institutions, and not-for-profit organisations such as Wikimedia Ireland could receive ongoing government support in establishing country-wide wiki-editing workshops.

Corporate Social Responsibility

Finally, as outlined in Section 1, Ireland is a hub for technological innovation. The growth witnessed in AI and NLP industries is mainly thanks to the positioning of EMEA headquarters for many global technology companies in Ireland. Yet, this innovation only serves the English-speaking public and economic community of Ireland. It would appear that, as part of a corporate social responsibility policy, investment into supporting the Irish language should now be given much more serious consideration.

Acknowledgements

This report has benefited from the insightful comments made by Maria Giagkou and Victoria Arranz. The author's gratitude goes to them as well as to those who provided invaluable suggestions to make this report as comprehensive as possible: Jane Dunne, Kevin Scannell, John Judge, Brian Ó Raghallaigh, Elaine Uí Dhonnchadha, Neasa Ní Chiaráin and Ailbhe Ní Chasaide. A special thanks to Sarah McGuinness, Emma Daly and Owen Gallagher for their work on the ELG data collection for Irish. Finally, thank you also to Shanna Ní Rabhartaigh for her usual level of excellence in the translation of the abstract.

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.

Mihael Arcan, Caoilfhionn Lane, Eoin Ó Droighneáin, and Paul Buitelaar. IRIS: English-Irish machine translation system. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 566–572, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020.

James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair, and Jennifer Foster. gaBERT - an Irish language model (preprint). *CoRR*, abs/2107.12930, 2021. URL https://arxiv.org/abs/2107.12930.

Aivars Berzins, Khalid Choukri, Maria Giagkou, Andrea Lösch, Helene Mazo, Stelios Piperidis, Mickaël Rigault, Eileen Schnur, Lilli Small, Josef van Genabith, Andrejs Vasiljevs, Andero Adamson, Dimitra Anastasiou, Natassa Avraamides-Haratsi, Núria Bel, Zoltán Bódi, António Branco, Gerhard Budin, Virginijus Dadurkevicius, Stijn de Smeytere, Hrístina Dobreva, Rickard Domeij, Jane Dunne, Kristine Eide, Claudia Foti, Maria Gavriilidou, Thibault Grouas, Normund Gruzitis, Jan Hajic, Barbara Heinisch, Verónique Hoste, Arne Jönsson, Fryni Kakoyianni-Doa, Sabine Kirchmeier, Svetla Koeva, Lucia Konturová, Jürgen Kotzian, Simon Krek, Gauti Kristmannsson, Kaisamari Kuhmonen, Krister Lindén, Teresa Lynn, Armands Magone, Maite Melero, Laura Mihailescu, Simonetta Montemagni, Micheál Õ Conaire, Jan Odijk, Maciej Ogrodniczuk, Pavel Pecina, Jon Arild Olsen, Bolette Sandford Pedersen, David Perez, Andras Repar, Ayla Rigouts Terryn, Eirikur Rögnvaldsson, Mike Rosner, Nancy Routzouni, Claudia Soria, Alexandra Soska, Donatienne Spiteri, Marko Tadic, Carole Tiberius, Dan Tufis, Andrius Utka, Paolo Vale, Piet van den Berg, Tamás Váradi, Kadri Vare, Andreas Witt, Francois Yvon, Janis Ziedins, and Miroslav Zumrik. *Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe – Why Language Data Matters: ELRC White Paper*. ELRC Consortium, 2 edition, 2019.

Lauren Cassidy, Teresa Lynn, James Barry, and Jennifer Foster. TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, May 2022.

Anna Čermáková, Jarmo Jantunen, Tommi Jauhiainen, John Kirk, Michal Křen, Marc Kupietz, and Elaine Uí Dhonnchadha. International comparable corpus: Challenges in building multilingual spoken and written comparable corpora. *Research in Corpus Linguistics*, 9(1):89–103, 2021. ISSN 2243-4712. doi: 10.32714/ricl.09.01.06.

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

Ronan Connolly. *TA needs survey of overseas Irish language learners*, pages 63–68. 12 2021. doi: 10.14705/rpnet.2021.54.1310.

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

Meghan Dowling, Lauren Cassidy, Eimear Maguire, Teresa Lynn, Ankit Srivastava, and John Judge. Tapadóir: developing a statistical machine translation engine and associated resources for Irish. In *4th Biennial Workshop on Less-Resourced Languages (LRC 2015)*, Poznan, Poland, 2015.

Meghan Dowling, Teresa Lynn, and Andy Way. Investigating backtranslation for the improvement of english-irish machine translation. *TEANGA, the Journal of the Irish Association for Applied Linguistics*, 26, 11 2019. doi: 10.35903/teanga.v26i0.88.

Meghan Dowling, Sheila Castilho, Joss Moorkens, Teresa Lynn, and Andy Way. A human evaluation of English-Irish statistical and neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 431–440, Lisboa, Portugal, November 2020. European Association for Machine Translation.

John Harris. *Irish in Primary Schools: Long-term National Trends in Achievement*. Stationery Office, 2006. ISBN 9780755773138.

Amelia C. Kelly, Harald Berthelsen, Nick Campbell, Ailbhe Ní Chasaide, and Christer Gobl. Corpus design techniques for Irish speech synthesis. In *China-Ireland International Conference on Information and Communications Technologies*, NUI Maynooth, Ireland, 2009.

Adam Kilgarriff, Michael Rundell, and Elaine Uí Dhonnchadha. Efficient corpus development for lexicography: Building the new corpus for ireland. *Language Resources and Evaluation*, 40:127–152, 02 2006. doi: 10.1007/s10579-006-9011-7.

Derek Lackaff and William Moner. Local languages, global networks: Mobile design for minority language users. In *Proceedings of the 34th Annual International Conference on the Design of Communication (SIGDOC '16)*, 09 2016. doi: 10.1145/2987592.2987612.

Seamus Lankford, Haithem Alfi, and Andy Way. Transformers for low-resource languages: Is féidir linn! In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 48–60, Virtual, August 2021. Association for Machine Translation in the Americas.

Teresa Lynn. *Irish Dependency Treebanking and Parsing*. PhD thesis, Dublin City University and Macquarie University, Sydney, 2016.

Teresa Lynn, Jennifer Foster, and Mark Dras. Working with a small dataset – semi-supervised dependency parsing for Irish. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–11, Seattle, Washington, USA, October 2013.

Mícheál John Ó Meachair. *The Creation and Complexity Analysis of a Corpus of Educational Materials in Irish (EduGA)*. PhD thesis, Trinity College Dublin School of Linguistic Speech and Computational Science, 2020.

Christopher Moseley. *The UNESCO atlas of the world's languages in danger*. World Oral Literature Project, 3 edition, 2012.

Michal Boleslav Měchura. Irish National Morphology Database: a high-accuracy open-source dataset of Irish words. In J. Judge, T. Lynn, M. Ward, and B. Ó Raghallaigh, editors, *Proceedings of the First Celtic Language Technology Workshop*, 2014.

Mairéad Nic Giolla Mhichíl, Theodore Lynn, and Pierangelo Rosati. Twitter and the Irish language, #Gaeilge – agents and activities: exploring a data set with micro-implementers in social media. *Journal of Multilingual and Multicultural Development*, 39, 03 2018. doi: 10.1080/01434632.2018.1450414.

Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Christoph Wendler, Harald Berthelsen, Andy Murphy, and Christer Gobl. The ABAIR initiative: Bringing spoken Irish into the digital space. In *Proceedings of INTERSPEECH 2017*, 08 2017. doi: 10.21437/Interspeech.2017-1407.

Neasa Ní Chiaráin and Ailbhe Chasaide. An scéalaí: autonomous learners harnessing speech and language technologies. In *Proceedings of the 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, pages 94–98, 09 2019. doi: 10.21437/SLaTE.2019-18.

Neasa Ní Chiaráin and Ailbhe Ní Chasaide. *The Potential of Text-to-Speech Synthesis in Computer-Assisted Language Learning*, pages 149–169. 01 2020. ISBN 9781799810995. doi: 10.4018/978-1-7998-1097-1.ch007.

Katie Ní Loingsigh. Towards a lexicon of Irish-language idioms. In *Proceedings of the Second Celtic Language Technology Workshop at JEP-TALN-RECITAL 2016*, Paris, France, 07 2016.

Katie Ní Loingsigh, Brian Ó Raghallaigh, and Gearóid Ó Cléircín. The design and development of Corpas na Gaeilge Comhaimseartha (Corpus of Contemporary Irish). In *Proceedings of the 9th International Corpus Linguistics Conference*, 07 2017.

Pádraig Ó Duibhir, Gabrielle NigUidhir, Seán Ó Cathalláin, and Jude Cosgrove. *An Analysis of Models of Provision For Irish-medium Education*. Foras na Gaeilge, 11 2017.

Conchúr Ó Giollagáin and Charlton Martin. *Nuashonrú ar an Staidéar Cuimsitheach Teangeolaíoch ar Úsáid na Gaeilge sa Ghaeltacht 2006–2011:: Príomhfhaisnéis na Limistéar Pleanála Teanga*. Údarás na Gaeltachta / Gaeltacht Development Authority, Ireland, May 2015.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92, 2014. doi: 10.1162/tacl_a_00167.

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.

Kevin Scannell. Statistical models for text normalization and machine translation. In *Proceedings of the First Celtic Language Technology Workshop*, pages 33–40, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. doi: 10.3115/v1/W14-4605.

Kevin P. Scannell. The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15, 2007.

Nancy Stenson. *Studies in Irish Syntax*. Ars linguistica. Narr, 1981. ISBN 9783878083580.

The Christian Brothers. *New Irish Grammar*. Dublin: C J Fallon, 1988.

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.

Elaine Uí Dhonnchadha. *Part-of-Speech Tagging and Partial Parsing for Irish using Finite-State Transducers and Constraint Grammar*. PhD thesis, Dublin City University, 2009.

Elaine Uí Dhonnchadha, Alessio Frenda, and Brian Vaughan. Issues in Designing a Corpus of Spoken Irish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1–6, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

Abigail Walsh, Teresa Lynn, and Jennifer Foster. Annotating verbal MWEs in Irish for the PARSEME shared task 1.2. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 58–65, online, December 2020. Association for Computational Linguistics.