EUROPEAN LANGUAGE EQUALITY

D1.34

Report on the Welsh Language

Authors	Delyth Prys, Gareth Watkins, Stefano Ghazzali
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	D1.34
Deliverable title	Report on the Welsh Language
Type	Report
Number of pages	29
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Authors	Delyth Prys, Gareth Watkins, Stefano Ghazzali
Reviewers	Teresa Lynn, Maria Eskevich
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

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

Contents

List of Figures

List of Tables

List of Acronyms

AHRC	Arts and Humanities Research Council
AI	Artificial Intelligence
BBC	British Broadcasting Corporation
BLARK	Basic Language Resource Kit
CL	Computational Linguistics
CV	Common Voice
CorCenCC	Corpws Cenedlaethol Cymraeg Cyfoes *(the National Corpus of Contemporary Welsh)*
DLE	Digital Language Equality
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRI	European Language Resource Infrastructure
ERDF	European Regional Development Fund
ESF	European Social Fund
ESRC	Economic and Social Research Council
EU	European Union
FAQ	Frequently Asked Question
GATE	General Architecture for Text Engineering
GPS	Global Positioning System
GPU	Graphics Processing Unit
HPC	High-Performance Computing
ICT	Information Communication Technology
IMLs	Indigenous, Minority and Lesser-used Languages
IP	Intellectual Property
IT	Information Technology
LR	Language Resource/Resources
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NER	Named-Entity Recognition

NHS	National Health Service
NLG	Natural Language Generation
NLP	Natural Language Processing
R&D	Research and Development
S4C	Sianel Pedwar Cymru *(the Welsh language television channel)*
SR	Speaker Recognition
TLD	Top Level Domain
UD	Universal Dependencies
UK	United Kingdom
UNESCO	United Nations Educational, Scientific and Cultural Organization
UTF8	Unicode Transformation Format 8
VSO	Verb-Subject-Object

Abstract

This study is part of a series that reports on the results of an investigation of the level of support European languages receive through technology, with a focus on the Welsh language. A thorough description of the Welsh language in the digital age is provided. This includes an overview of the status of Welsh and English in Wales. The use of Welsh in education and the efforts made to standardise technical terminology for the Welsh education system, along with a summary of the Welsh writing system and the typology of Welsh, a member of the Brythonic Celtic group, are presented. The use of Welsh online, along with Welsh people's level of access to technology is covered. The scope of this report in respect of Language Technology (LT) is then laid out, and a general definition of LT provided. The LT support available to the Welsh language is discussed. This discussion includes the availability of corpora and the suitability of these corpora for reuse in respect of licencing and IP issues, together with recent improvements. The digitisation of Welsh dictionaries along with the availability of other resources such as spellcheckers, grammar checkers and specialised lexica are discussed, and recent developments in respect of creation of acoustic and language models for Welsh are outlined. The many and varied tools available to the Welsh language are then introduced, resources such as NLP tools, virtual personal assistants, chatbots, speech technology, and translation technology. Some of these tools have been developed specifically for Welsh, while others are part of international frameworks. Key projects, initiatives and stakeholders are then identified. Responsibility for LT and Artificial Intelligence (AI) in the service of the Welsh language resides with the Welsh Government, however the relevance of the development of the UK's National AI Strategy is also considered. The importance of LT's place in Wales in respect of language revitalisation and economic regeneration is underlined, along with the key strategies related to these areas. Much effort has been put into developing infrastructure relating to the Welsh language in recent years and this infrastructure is constantly growing. The significance of the Welsh Government's Welsh Language Technology Action Plan together with further actions which followed in the wake of this plan are underlined. While the UK LT industry is mostly focused on the needs of the English language, Welsh language LT provision is mainly driven forward by the higher education sector. Wales does have vibrant creative technology, media and translation sectors which make use of the government funded opensource resources created by universities. Welsh language LT provision is compared to other European languages, based on an empirical investigation of the resources available in the European Language Grid catalogue. In conclusion we find that Welsh LT has made considerable progress in the digital sphere in the last few years, however further development and investment is needed. Gaps in provision are identified, such as bilingual models needed for bilingual language communities, and suggestions made illustrating how new Welsh language tools and resources can fill those gaps. Finally, the need for continuing development and investment is emphasised, so that minoritised languages such as Welsh do not fall behind as new digital services and products come to market.

Crynodeb

Mae'r astudiaeth hon yn rhan o gyfres sy'n adrodd ar ganlyniadau ymchwil i'r lefel o gefnogaeth gaiff ieithoedd Ewrop drwy dechnoleg, gan ganolbwyntio ar y Gymraeg. Rhoddir dadansoddiad manwl o'r Gymraeg yn yr oes ddigidol. Mae hyn yn cynnwys trosolwg o safle'r Gymraeg a'r Saesneg yng Nghymru, lle mae statws swyddogol i'r Gymraeg a lle mae deddfwriaeth yn datgan na ddylid trin y Gymraeg yn llai ffafriol na'r Saesneg.

Cyflwynir y defnydd o'r Gymraeg mewn addysg, lle mae dysgu'r Gymraeg yn orfodol i blant rhwng 5 – 16 oed o fewn y Cwricwlwm Cenedlaethol. Mae dewis o ysgolion cyfrwng Cym-

raeg, cyfrwng Saesneg a dwyieithog ar gael, a cheir dysgu trwy gyfrwng y Gymraeg ar lefel chweched dosbarth ac addysg bellach. Ehangwyd addysg cyfrwng Cymraeg ar lefel prifysgol yn ddiweddar hefyd, drwy'r Coleg Cymraeg Cenedlaethol sydd â phresenoldeb ym mhob un o brifysgolion Cymru. Law yn llaw ag ehangu addysg cyfrwng Cymraeg, cafwyd ymdrechion i safoni termau technegol ar gyfer system addysg Cymru, sydd wedi creu adnoddau cyfoethog terminolegol a gaiff eu storio mewn cronfeydd data electronig.

Ceir 29 llythyren yn y wyddor Gymraeg, yn cynnwys y deugraffau ch, dd, ff, ng, ll, ph, rh a th. Ni ddefnyddir y llythrennau Saesneg v, x a z yn Gymraeg, ond cânt eu cynnwys gyda'r wyddor Gymraeg at ddibenion cyfrifiadurol gan eu bod yn digwydd mewn enwau endidau megis enwau personol ac enwau lleoedd. Defnyddir nifer o nodau acennog yn y Gymraeg, yn arbennig yr acen grom, acenion dyrchafedig a disgynedig a'r didolnod. Roedd ysgrifennu'r nodau Ŵ, ŵ, Ŷ ac ŷ yn broblemus iawn yn y cyfryngau digidol cyn dyfodiad safon UTF8 ond bellach mae wedi'i ddatrys os defnyddir y safon honno.

Mae'r Gymraeg yn perthyn i'r grŵp ieithoedd Brythonig Celtaidd, ac yn dilyn trefn geiriau Berf-Goddrych-Gwrthrych. Mae hefyd yn hynod oherwydd y treigliadau geir ar ddechrau geiriau, a oedd yn cynnig her arbennig wrth ddechrau trin y Gymraeg yn gyfrifiadurol, ac wedi golygu y bu'n rhaid datblygu offer iaith megis lemateiddwyr i ddelio â nhw. Ceir nifer o dafodieithoedd daearyddol yn y Gymraeg, a hefyd amrywiaeth fawr o ran cyweiriau iaith, o'r cywair tra ffurfiol i gyweiriau mwy anffurfiol, gan gynnwys cyweiriau sathredig a chyfnewid cod gyda'r Saesneg.

Ceir lefel uchel o lythrennedd digidol yng Nghymru, ac yn 2019 roedd gan 90% o'r boblogaeth fynediad at y rhyngrwyd. Trafodir y defnydd o'r Gymraeg ar-lein, ynghyd â lefel mynediad pobl Cymru i dechnoleg yn y ddogfen hon. Mae diffyg offer technoleg hygyrch yn y Gymraeg yn effeithio'n andwyol ar allu'r cyhoedd i ddefnyddio'r Gymraeg ar y we ac yn y cyfryngau cymdeithasol, yn enwedig os ydynt yn ddefnyddwyr llai hyderus eu Cymraeg, fel sy'n digwydd yn aml mewn cymunedau ieithoedd lleiafrifedig.

Trafodir y gefnogaeth dechnoleg iaith sydd ar gael i'r Gymraeg. Sonnir am argaeledd corpora ac addasrwydd y corpora hyn i gael eu hail ddefnyddio o ran materion trwyddedu a hawlfraint, ynghyd â gwelliannau diweddar. Corpws testun Cysill Ar-lein yw'r corpws mwyaf o ddigon o ran maint ar gyfer y Gymraeg, yn cynnwys dros 400 miliwn tocyn ac yn dal i dyfu. Corpws CorCenCC yw'r corpws mwyaf sydd wedi'i anodi, gyda dros 11 miliwn tocyn. Mae corpws anodedig Siarad o sgyrsiau dwyieithog hefyd yn werth ei grybwyll. Ceir y casgliad mwyaf o gorpora gan Brifysgol Bangor gyda thros 700,000,000 tocyn, ond nid yw'r cyfan ar gael yn agored, ac mae materion hawlfraint a modelau trwyddedu yn bwnc llosg wrth geisio rhannu data yn agored.

Mae nifer o eiriaduron Cymraeg ar gael ar ffurf ddigidol, yn ogystal â gwirwyr sillafu a gramadeg, a lecsica arbenigol. Mae Hunspell yn wirydd sillafu cod agored sydd ar gael ar gyfer nifer o ieithoedd, gan gynnwys y Gymraeg, a Cysill yn wirydd sillafu a gramadeg poblogaidd. Yn fwy diweddar crëwyd modelau acwstig a modelau iaith Cymraeg. Ceir adnoddau ac offer ar gyfer prosesu iaith naturiol, cynorthwywyr personol rhithiol, sgwrsfotiaid, technoleg lleferydd a thechnoleg cyfieithu. Datblygwyd rhai o'r rhain yn benodol ar gyfer y Gymraeg, tra bo eraill yn rhan o fframweithiau rhyngwladol, megis un Common Voice gan Mozilla a Wicipedia Cymraeg. Mae Trawsgrifiwr yn wasanaeth newydd yn y Gymraeg sy'n galluogi trawsgrifio lleferydd yn destun, a Macsen yw'r ap cynorthwyydd personol Cymraeg cyntaf sy'n cael ei ddatblygu gyda set o sgiliau y mae modd i ddatblygwyr eraill ychwanegu atynt. Ceir nifer o raglenni cyfieithu peirianyddol Cymraeg, gydag un benodol ar gyfer y parth Iechyd a Gofal yn cael ei ddatblygu ar hyn o bryd. Ceir hefyd lwyfan cyhoeddus cofion.techiaith.cymru lle gall cyrff cyhoeddus rannu eu cofion cyfieithu ymhlith ei gilydd.

Mae cyfrifoldeb am Dechnoleg Iaith a Deallusrwydd Artiffisial yn gorwedd gyda Llywodraeth Cymru, ond mae Strategaeth Genedlaethol Deallusrwydd Artiffisial y DU hefyd yn berthnasol. Tanlinellir pwysigrwydd lle technoleg iaith yng Nghymru o ran adfer iaith ac adfywio economaidd, ynghyd â strategaethau allweddol yn perthyn i'r meysydd hyn. Prif strateg-

aeth Llywodraeth Cymru ynglŷn â'r Gymraeg yw 'Cymraeg 2050: Miliwn o siaradwyr Cymraeg' sy'n pwysleisio lle technoleg iaith i gefnogi addysg, Cymraeg yn y gweithle a defnydd cymdeithasol o'r Gymraeg. Manylwyd ar hyn yn 'Cynllun Gweithredu Technoleg Gymraeg' Llywodraeth Cymru, sy'n nodi tri maes blaenoriaeth, sef Technoleg Lleferydd Cymraeg, Cyfieithu â Chymorth Cyfrifiadur, a Deallusrwydd Artiffisial Sgwrsiol. Ymdrechwyd yn galed i wella'r isadeiledd sydd yn cynnal y Gymraeg yn y blynyddoedd diweddar, ac mae'r isadeiledd hynny yn tyfu'n gyson.

Tra bo diwydiant technoleg iaith y DU yn ffocysu yn bennaf ar anghenion yr iaith Saesneg, prif symbolwr darpariaeth technoleg iaith Gymraeg yw'r sector addysg uwch lle mae'r rhan fwyaf o ymchwil yn y maes yn digwydd. Mae gan Gymru sectorau diwydiannau creadigol a'r cyfryngau a chyfieithu sydd yn gwneud defnydd o adnoddau cod agored a ariannwyd gan y llywodraeth ac a grëwyd gan brifysgolion. Drwy hyn gobeithir bod datblygiad technoleg Iaith Gymraeg hefyd yn cyfrannu at adfywiad economaidd y wlad.

Ceir Rhwydwaith Technolegau Iaith Genedlaethol i Gymru, yn cynnwys aelodau o brifysgolion a byd diwydiant. Ceir grŵp ymchwil rhyngwladol Technolegau Iaith Geltaidd, sydd â chynrychiolaeth o'r holl ieithoedd a gwledydd Celtaidd. Ceir hefyd adnodd uwchgyfrifiadura cenedlaethol, Supercomputing Wales, sy'n fenter ar y cyd rhwng prifysgolion Caerdydd, Abertawe, Bangor ac Aberystwyth. Mae cymhwysedd digidol bellach yn fframwaith trawsgwricwlaidd sy'n ceisio integreiddio sgiliau digidol ar draws nifer o feysydd gwahanol. Ymdrechir i ehangu gwyddor technoleg iaith ar lefel prifysgol hefyd, gyda gradd Meistr mewn Technoleg Iaith yn weithredol ym Mhrifysgol Bangor ers 2020.

Cymharir darpariaeth technoleg iaith Gymraeg ochr yn ochr ag ieithoedd Ewropeaidd eraill, yn seiliedig ar astudiaeth empirig o'r adnoddau sydd ar gael yng nghatalog Grid Ieithoedd Ewrop (ELG). Yn sgil y cynnydd mewn ymchwil ym maes technoleg iaith ar gyfer y Gymraeg yn y blynyddoedd diweddar, ac yn arbennig y project 'Technoleg a'r Gymraeg' a gychwynnwyd ym Mhrifysgol Bangor yn 2020, ac a ariannwyd gan Lywodraeth Cymru, cynyddodd y nifer o adnoddau, offer, gwasanaethau a chynnyrch Cymraeg. Rhestrir y cyfan y llwyddwyd i gael gwybodaeth amdanynt yng nghatalog ELG.

Eto i gyd, ochr yn ochr â'r cynnydd a fu yn y ddarpariaeth ar gyfer ieithoedd mawr Ewrop, nid yw'r ddarpariaeth ar gyfer y Gymraeg yn dod yn agos at gau'r bwlch rhyngddi a'r ieithoedd hynny. Yn sgil ymadawiad y DU â'r Undeb Ewropeaidd, nid oes gan y Gymraeg bellach bresenoldeb yno, ac mae hynny'n golled fawr iddi. Y mae'r Gymraeg serch hynny yn rhan o waddol ieithyddol gyfoethog Ewrop a'i dymuniad yw chwarae rhan lawn mewn datblygu atebion arloesol i'n problemau cyffredin, a sicrhau cydraddoldeb digidol i holl ieithoedd Ewrop.

1 Introduction

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

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

[1] The results of this data collection procedure have been integrated into the European Language Grid so that they

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

2 The Welsh Language in the Digital Age

2.1 General Facts

Status of the Welsh language

The Welsh language is mainly spoken in Wales, together with a small population in the province of Chubut, Argentina following historical emigration from Wales. In some regions of north and west Wales as much as 76.4% of the population speak Welsh (StatsWales, 2021). Although a smaller percentage speak Welsh in south east Wales, the number of Welsh speakers is higher there due to the overall population density (ibid).

Welsh is classed as a minority language under the European Charter for Regional or Minority Languages (Prys, 2006), and is considered "vulnerable" by UNESCO's Atlas of the World's Languages in Danger (Moseley, 2010). The Welsh Language (Wales) Measure 2011 states that "The Welsh language has official status in Wales" (National Assembly for Wales, 2011), and "makes provision for promoting and facilitating the use of the Welsh language and treating Welsh no less favourably than English" (Welsh Language Commissioner, 2021a). Incidentally, this makes English also an official language in Wales, co-official with Welsh, the only part of the UK where it has explicit official status (Mac Síthigh, 2018). The 2011 measure also created the role of Welsh Language Commissioner whose remit is to promote the use of Welsh, and can impose Welsh language standards on organisations, with power to enforce those standards. The Welsh Language Commissioner's website[2] includes a wealth of statutory and advisory guidelines, including guidance on how Information Technology (IT) systems should provide Welsh language and bilingual capabilities.[3]

In 2012, the results of the 2011 Census were released. These results showed a decline of 20,000 in the number of Welsh speakers compared to the 2001 Census, reporting that there were 562,000 (or 19% of the population of Wales) Welsh speakers in Wales compared to the 582,000 (or 20.8% of the population of Wales) reported in the 2001 Census. The results of the 2021 decennial census are due to be released in late spring 2022.[4] It is hoped that the figures will show an increase in the number of Welsh speakers in line with the Welsh Government's strategy to achieve a million Welsh speakers by 2050.[5]

The 2011 Census results were "a catalyst for much activity", including formulation of the Welsh Governments Welsh language strategy, titled "Cymraeg 2050: A million Welsh speakers" (Welsh Language Commissioner, 2021b). In this strategy, published in 2017, the Welsh Government document their ambition of seeing a million Welsh speakers by 2050. They recognise the importance of technology, and see investment in the development of technology as key to this ambition (Welsh Government, 2017a).

can be discovered, browsed and further investigated by means of comparative visualisations across languages.

2 https://www.welshlanguagecommissioner.wales

3 https://www.welshlanguagecommissioner.wales/media/2efhxwjn/technoleg-gwefannau-a-meddalwedd-technology-websites-and-software.pdf

4 https://www.ons.gov.uk/census/censustransformationprogramme/censusnews

5 https://research.senedd.wales/research-articles/is-a-million-welsh-speakers-by-2050-achievable/

Use in Education

The teaching of Welsh, either as a first or second language, is compulsory for children aged 5 – 16 in Wales, within the National Curriculum (Welsh Government, 2017b). In the academic year 2020-2021 there were 78,000 pupils attending Welsh medium schools, 347,000 attending English medium schools, and 32,000 attending dual stream or bilingual schools, where a percentage of lessons are taught through the medium of Welsh alone.[6] Welsh as a subject and medium of education continues to be offered at A level (pupils aged 16-19) and further education (occupational training) level. Students can study around ‘1,000 courses partly or entirely through the medium of Welsh’ at university level in Welsh universities through the Coleg Cymraeg Cenedlaethol, a virtual Welsh-medium college with presence in all the Welsh universities.[7] Great efforts have been made to standardise technical terminology for the Welsh education system, with bilingual terminology dictionaries continuously updated for the school national curriculum and university teaching and research (Andrews and Prys, 2016). These are stored in electronic master databases enabling their publication in many different formats to aid dissemination, including online. The Ap Geiriaduron dictionary app which also contains many of these dictionaries, had been downloaded 273,670 times to date (39,670 for the Android version and 234,000 for the iOS version),[8] making it one of the most popular Welsh language apps, greatly valued and used by Welsh learners, as well as within Welsh-medium education.

Writing system

Welsh is written with the Latin writing script, adapted for Welsh. As is illustrated in Table 1, the Welsh alphabet contains 29 letters, including the letter ‘j’ borrowed from English to represent the borrowed /dʒ/ consonant phoneme. Another borrowed sound from English is the consonant phoneme /tʃ/, but as this is usually written ‘ts’ it does not add another letter to the Welsh alphabet. The letters v, x and z found in English are not used in Welsh, however they are included with the Welsh alphabet for computer use as they often appear in named entities, including personal names, foreign placenames and names of international products and organisations. Eight letters in the Welsh alphabet are digraphs, namely ch, dd, ff, ng, ll, ph, rh, th. This means that they are treated as single letters for purposes such as crossword puzzles, and for sorting purposes. Thus, all words beginning with ‘ll’, for example, (representing the [ɬ] sound), would appear in a dictionary after all words beginning with ‘l’ and not in the middle of the sort order for ‘l’.

a	b	c	ch	d	dd
e	f	ff	g	ng	h
i	j	l	ll	m	n
o	p	ph	r	rh	s
t	th	u	w	y	

Table 1: The Welsh Alphabet

Accented characters are common over vowels in Welsh, most usually the circumflex accent, the acute and grave accents and diaresis mark, all occurring over vowels to help denote pronunciation. The Ŵ, ŵ, Ŷ and ŷ characters were especially problematic for writing Welsh in digital media before the advent of the UTF8 standard as they are rare in major European

[6] https://statswales.gov.wales/Catalogue/Education-and-Skills/Schools-and-Teachers/Schools-Census/Pupil-Level-Annual-School-Census/Welsh-Language/pupils-by-localauthorityregion-welshmediumtype

[7] https://www.colegcymraeg.ac.uk/en/study/mediumofwelsh/

[8] Figures correct as of December 2021

languages and were not covered by earlier conventions.[9] Accented vowels are not distinguished from non-accented vowels for the purpose of sorting in alphabetical order.

Welsh has a long tradition of a standard written language, going at least as far back as the first complete translation of the Bible into Welsh in 1588. The modern orthography was standardised in 1928, with the publication of the Orthography Panel's recommendations (University of Wales, 1928) and further minor amendments in 1987 (Lewis, 2018). In 2021 the Welsh Government established a new Welsh Orthography Panel, which will attempt to resolve minor inconsistencies in orthography, notably for accented characters and hyphenation.

Welsh spelling is usually considered to be fairly regular and phonetic, but some features, such as the doubling of the 'n' and 'r' consonants, accented characters, and confusion over writing 'u', 'y' and 'i' (especially by people speaking dialects where the /ɨ/ and /ɨː/ sounds are not present) still present orthographic challenges to the uninitiated.

Typology of Welsh

Welsh belongs to the insular Celtic branch of Indo-European languages, which is further subdivided into the Brythonic Celtic group (Welsh, Cornish and Breton), and the Goedelic Celtic group (Irish, Scottish Gaelic and Manx). The Brythonic group is also called the 'P' Celtic group, and the Goedelic group the 'Q' Celtic group, following the substitution of p- or the retention of the original Indo-European qu- respectively in both groups.

In common with other Celtic languages, Welsh is verb initial, following a VSO (verb-subject-object) order. It also shares the Celtic peculiarity of consonant mutations at the beginning of words, which posed a particular challenge in the early days of developing computational tools for Welsh. Alongside the standard written language, Welsh has a continuum of other registers, with colloquial or informal registers differing markedly from the standard written form, and with many regional varieties of accents and dialects. The main dialect differences are between north and south, with northern dialects having the additional /ɨ/ and /ɨː/ vowels no longer present in southern Welsh dialects.

Welsh has two methods of verb formation, with formal registers, the standard written form and some dialects using the concise forms, but colloquial registers inclined to use periphrastic forms, using auxiliary verbs, which is increasingly affecting the standard written language. Another marked difference between formal and colloquial language registers is the use of code switching with English in more informal registers (Prys, 2016). This is often a mark of bilingual societies, with youth culture in particular making creative use of their fluency in two languages to promote their own identity. Increased use of social media in Welsh has highlighted this trend and presents another challenge for language tools attempting to deal with Welsh and English in bilingual contexts.

Stress in Welsh is fairly regular, falling on the penultimate syllable in polysyllabic words with some exceptions which are usually marked through the use of accented characters. Some Welsh sounds are rare in other European languages, and do not occur at all in English. These include the voiceless alveolar lateral fricative [ɬ] and several voiceless sonorants mainly resulting from nasal mutations. These present a problem for automatic speech synthesis and recognition programs in other languages as well as Welsh. Inability to recognise and pronounce Welsh named entities such as personal names and placenames containing these sounds result in lower accuracy in speech-dependent programs, such as failure, for example, to retrieve music by the Welsh singer Lleuwen from music playlist or to find the town of Llanelli on a GPS system. These particular Welsh sounds are dealt with in the letter-to-sound rules published as part of the suite of resources to aid the development of Welsh speech technology.[10]

[9] https://sites.psu.edu/symbolcodes/languages/europe/welsh/
[10] https://github.com/techiaith/welsh-lts

A comprehensive up-to-date description of Welsh for LT purposes may be found in (Cunliffe et al., 2021).

2.2 Welsh in the Digital Sphere

In general the people of Wales are internet literate. The Welsh Government (2021) note that 90% of the 2019/2020 National Survey for Wales' respondents used the internet, that 73% of those people demonstrated the ability to use all five of the basic digital skills specified in the digital inclusion framework,[11] and that 94% of internet using respondents sent a message via email or instant messaging, or posted on social media in the three months prior to being polled. Song et al. (2020) calculate that 8 in 10 of the people of Wales use social media in one form or another. However neither the Welsh Government (2021) nor Song et al. (2020) specify the language used when accessing the internet or social media.

The results of research conducted by Beufort Research, commissioned by BBC Cymru Wales, S4C (the Welsh language television channel) and the Welsh Government in 2013 "show that much higher proportions of Welsh speakers are communicating online in English than in Welsh" (BBC Cymru Wales and S4C and Welsh Government, 2013). Similar findings were reported by the Welsh Government (2015). This is possibly to be expected in a country where more than one language has official status but not all residents are bilingual. As noted by BBC Cymru Wales and S4C and Welsh Government (2013) "In certain social media environments, in particular Facebook, some participants were not consciously thinking of whether or not they should post in Welsh because the range of friends included people who could not speak the language". A lack of accessible language tools for Welsh, including proofing tools, exacerbates the problem, especially for less confident writers, which may form a significant proportion of Welsh speakers, in common with other minoritised language communities. It's possible that the percentage of people using Welsh in social media has increased in the 7-9 years since these reports were published, as the quality of Machine Translation (MT) and other language tools has significantly increased in recent years, allowing friends and followers of social media accounts or recipients of e-mails to understand Welsh language content even if they are less confident in their own use of Welsh or are unable to understand the Welsh language. This underlines the value in investment in Language Technology (LT) to a minoritised language such as Welsh.

According to nTLDstats,[12] which 'collates [data] from root zone reports on domain registrations that are logged with ICANN',[13] on 04/11/21, there were 8,108 websites registered with the .cymru Top Level Domain (TLD), and 13,955 websites registered with the .wales TLD. No mention is made of the language of the content of sites bearing the .wales and .cymru TLDs, and no such figures currently exist.

The Welsh Government suggest that 87% of households have access to the internet (Welsh Government, 2019). Ofcom (2020)[14] however state that 94% of homes in Wales are able to access superfast broadband, and 97% have a 'decent' fixed broadband connection. Conversely, 1.2% have no access to a 'decent' broadband connection whatsoever. In respect of mobile internet access, there is 60% 4G geographic coverage in Wales from all four mobile network operators, and 90% geographic coverage from at least one operator. Conversely, Ofcom 'estimate that 0.6% (9,000) of premises cannot access either a decent fixed broadband service or get good 4G coverage indoors, with almost all of these in rural areas.' 5G continues to be rolled out in Wales, with 'current 5G deployments for consumers largely focusing on delivering mobile broadband, particularly in areas of existing high demand.'

[11] https://gov.wales/sites/default/files/publications/2019-05/digital-inclusion-framework-report-and-forward-look.pdf
[12] https://ntldstats.com/tld
[13] https://www.nominet.uk/cymru-wales-19000-domain-names-for-wales/
[14] The UK's communication regulator

According to Cunliffe et al. (2021), "on the Digital Language Vitality Scale (Ceberio et al., 2018), Welsh is 'Developing', arguably tending towards 'Vital' in some aspects".[15]

3 What is Language Technology?

Natural language[16] 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

[15] Note that a language is "developing" where "The language is visible on the Internet and is used over communication and social media, although frequency may still be occasional. Some digital media and services may be available, as well as a Wikipedia; basic (electronic) language resources exist, and there might be evidence of more advanced ones. At least one among the social media and the operating systems used by the speakers' community might be localised. An online machine translation service or tool might be available, for one language pair at least", while a language is "vital" where "The language is highly present on the Internet, and is used regularly for e-communication and on social media, some of which may have a localised interface. There is a considerable variety of digital media available. Language resources are widely available. Wikipedia projects are big and actively used/participated. The language can be used in all digital domains. Most used operating systems and general purpose software are localised in the language. There is evidence of machine translation tools/services".

[16] 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).

modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained with vast amounts of data, be it text, audio or multimodal. The success in these areas of AI has been possible because of the conjunction of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual data), 3) increase in high performance computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches.

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

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

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

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[17]

[17] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is

4 Language Technology support for the Welsh Language

4.1 Language Data

Corpus Resources

The availability of both text and speech corpora for Welsh has much improved in recent years. Monolingual text corpora, bilingual or multilingual corpora, and speech corpora, all of contemporary Welsh, and mainly in the standard or neutral language register, have been curated by various Welsh language stakeholders. The largest of these by far is the Cysill Arlein text corpus which is a monitor corpus which has reached 400 million tokens in size.[18] It is collated through the novel method of collecting texts inputted into the free online version of Bangor University's Welsh spelling and grammar checker, Cysill. A fuller description of this corpus may be found in (Prys and Watkins, 2021). The CorCenCC corpus (Knight et al., 2020) is the largest annotated, balanced general corpus to date. The corpus is annotated in respect of part of speech and semantic meaning and was the fruit of a project led by Cardiff University with participation from Swansea, Leicester and Bangor Universities. The CorCenCC corpus contains over 11 million tokens from written, spoken and electronic (online, digital texts) Welsh language sources, taken from a representative range of genres, language varieties (regional and social) and contexts. The Siarad corpus of bilingual (Welsh and English) conversations is another annotated corpus of note, resulting from a project at Bangor University's Bilingualism Centre. The Language Technologies Unit at Bangor University holds the largest collection of Welsh language corpora, totalling over 700,000,000 tokens, but its use is restricted to research and development within the university only, and for the creation of language models. Intellectual Property (IP) and licencing issues are of utmost concern when assessing the suitability of these corpora for use and reuse, and efforts are under way to promote CC-0 style open, permissive licences so that such data can be reused in the formation of tools and services without restriction.

A parallel corpus has been created using the bilingual record of proceedings of the National Assembly for Wales. It contains data from 1999-2003 and 2007-2010. Bangor University's Language Technology Unit has created an online tool to search this corpus. Neither monolingual nor bilingual domain specific corpora have been curated thus far. Such resources would be desirable as they would feed the development of domain specific tools.

Speech corpora for Welsh are a more recent development, and have been produced largely for use in speech technology projects. Crowdsourcing has been successfully used to gather large speech corpora of recorded prompts, initially through the Paldaruo app, subsequently using Mozilla's Common Voice system, which currently holds 143 hours of recorded speech donated by over 1,600 volunteers. Recordings of voice talents (male and female with different regional accents) which were collected specifically for building synthetic voices are currently in the process of being released under the CC-0 licence. These may also be considered as speech corpora, available for reuse in other environments.

Lexical/Conceptual Resources

The major, traditional, paper-based dictionaries have been digitised and ongoing lexical work now occurs natively in a digital environment. The comprehensive historical dictio-

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

[18] Figures correct as of December 2021

nary of Welsh[19] was digitised and made available online in 2014 and since 2016 has existed in app form. The main, contemporary English to Welsh dictionary was digitised in 2012 and is now available online.[20] In contrast to traditional descriptive dictionaries, terminology work in Welsh has for many years been concept based, held in computer databases and published in multiple formats. These resources have been recycled for use in lexicons for various purposes, including spelling and grammar checkers.

Bangor University's Welsh-language Lexicon, published in 2020 and updated in 2021 is a comprehensive wordlist of Welsh words and wordforms along with their respective grammar information, containing 828,746 entries. This was built over a number of years in an iterative process with resources such as the Hunspell spelling checker and Cysill spelling and grammar checker both contributing to it and drawing from it in turn. More specialised lexicons include stoplists for various purposes, while a thesaurus is available as part of the Cysill spelling and grammar checker. A Welsh WordNet has also been developed and released under the FreeBSD licence by a multi-disciplinary group at Cardiff University.

Models and Grammars

The creation of acoustic and language models for Welsh is a more recent development. Some of these models were part of multilingual sets which are of variable quality. Those developed specifically for Welsh tend to be of higher quality. Some have been developed for specific purposes such as helping to build a transcription system, and released under permissive open licences so that they may be reused by others without restriction. A Welsh language Part of Speech tagging model has been developed for spaCy, thus unlocking the potential to perform many other NLP tasks on Welsh language text. Further models are being developed to add NER and anonymisation components to the Welsh language spaCy pipeline.

4.2 Language Technologies and Tools

Tools and services developed for Welsh are many and varied. They include NLP tools for text analysis, anonymisation, and information extraction. Some tools are developed specifically for Welsh, while others are part of international frameworks such as the Welsh Natural Language Toolkit for the GATE framework (Cunliffe et al., 2021) or Stanford University's Stanza project.[21] which uses Universal Dependencies (UD) data. An alternative UD tool is currently under development by Bangor University's Language Technology Unit. The first text summarisation tool for Welsh is also under development.[22]

A Welsh virtual personal assistant called Macsen[23] exists as a free downloadable product, with a limited but expanding set of skills that may be added to by other developers. BOBi,[24] a bilingual FAQ generation system, has been developed by Cardiff City Council and is being used to support the citizens of Cardiff. Trawsgrifiwr,[25] the first Welsh speech to text transciber, has also recently been published again with the possibility of integrating it into other products and services. The Vocab 'mouse over' dictionary for websites is a useful service, enabling non-Welsh speakers and learners to access Welsh language websites.

In terms of Text-To-Speech, some synthetic voices have been created for Welsh using older diphone technology, with newer, more natural sounding unit selection voices now becoming

[19] https://www.welsh-dictionary.ac.uk
[20] https://geiriaduracademi.org
[21] https://stanfordnlp.github.io/stanza/
[22] https://corcencc.org/resources
[23] http://techiaith.cymru/packages/macsen/?lang=en
[24] https://www.cardiff.gov.uk/ENG/Home/Contact-us/Chatbot/Pages/default.aspx
[25] https://github.com/techiaith/trawsgrifiwr-arlein

available, such as the commercial Ivona voices (since acquired by Amazon). A new generation of synthetic voices, to be released under open licences, are under development also. A voice banking initiative called Lleisiwr, a joint venture between Bangor University and NHS Wales, has been created for bilingual Welsh/English speakers about to lose their natural speech capabilities, and is one of the most innovative services established to date.

In terms of translation technology, a dedicated commercial Welsh – English translation system exists and MT for Welsh is offered by some major companies such as Google and Microsoft Bing. Apertium is available for Welsh as an open source, rule-based MT platform, and Moses, another open source system, has been much used to develop statistical based MT for Welsh (Jones et al., 2019). These are now used by some translation companies in Wales to help their commercial translation services (Prys, 2021). Newer neural net frameworks are now being used, including MarianNMT, and the first domain-specific MT engine for Health will be launched in the coming months. Of particular note is the service provided by the Open Translation Memories website[26] where Welsh/English translation memories can be uploaded, downloaded, and shared. It is an emulation of the European ELRI project,[27] a project Welsh could not be included in since it was not an official EU language.

4.3 Projects, Initiatives, Stakeholders

National programme for LT/AI

The Welsh language is a devolved issue in the UK. Consequently responsibility for LT and AI in the service of the Welsh language resides with the Welsh Government, and not in London. In September 2021, the UK's National Artificial Intelligence (AI) Strategy was launched with the aim of helping the UK 'strengthen its position as a global science superpower and seize the potential of modern technology to improve people's lives and solve global challenges such as climate change and public health'.[28] While the document emphasises the importance of AI for economic development, and notes the need to develop AI standards, ethics and infrastructure for the UK, there is scant mention of language-orientated AI, and other than "levelling up digital prosperity within the UK" no reference is made to UK nations and regions, or their needs for LT in languages other than English.

In Wales, LT is seen primarily as a vehicle for language revitalisation, but together with AI is also part of the economic regeneration agenda. The current Welsh Government's Welsh language strategy "Cymraeg 2050: A million Welsh speakers", states that "We must ensure that high-quality Welsh language technology becomes available during the early stages of this strategy to support education, workplaces and social use of Welsh" (Welsh Government, 2017a). This was further elaborated in the Government's Welsh Language Technology Action Plan (Welsh Government, 2018) aiming "to plan technological developments to ensure that the Welsh language can be used in a wide variety of contexts, be that by using voice, keyboard or other means of human-computer interaction." Three main areas were identified for priority action, namely:

1. Welsh Language Speech Technology
2. Computer-assisted translation
3. Conversational Artificial Intelligence

[26] https://cofion.techiaith.cymru
[27] http://www.elri-project.eu
[28] https://www.gov.uk/government/news/new-ten-year-plan-to-make-britain-a-global-ai-superpower

The other strategy of note is the Welsh Government's "Prosperity for All: economic action plan".[29] This has R&D, Automation and Digitalisation as one of its 5 calls to action, identifying universities in Wales as having an important role to play in delivering this action.

The British and Irish Council, with membership comprising of representatives from the Irish Government; UK Government; Scottish Government; Northern Ireland Executive; Welsh Government; Isle of Man Government; Government of Jersey and Government of Guernsey, has a special interest in indigenous, minority and lesser-used languages (IMLs). This sector work group is led by the Welsh Government. Their current workplan for the 9 IMLs (Welsh, Irish, Ulster Scots, Gaelic, Scots, Manx, Cornish, Jèrriais, Guernésiais) includes a focus on Infrastructure / Technology / Economic Impact, the other two focuses being Social Use of Language (Broadcasting / Social Media) and Early Years. In 2015 the British and Irish Council held a special conference in Dublin on 'Promoting Our Languages Through Technology' (British-Irish Council, 2017) which laid the ground for a common understanding of LT requirements for the IMLs in the British Isles and Ireland, including Welsh.

National research infrastructures/e-infrastructures supporting or relating to language and LT

Wales has three national portals of relevance: the National Language Technologies Portal,[30] National Terminology Portal[31] and National Corpora Portal[32] which provide 'one stop shops' for information and tools and resources for these fields.

Wales also has a National Welsh Language Technologies Network[33] which currently has 117 members drawn from across academia, industry, public organisations and voluntary bodies. The main hubs for LT research in Wales are Bangor University where there is a dedicated LT Research Unit, and the universities of Cardiff, Swansea and South Wales. A dedicated annual academic symposium for LT in Wales brings these researchers together, with their proceedings published in parallel in both Welsh and English.[34] Wales is also part of the Celtic Language Technologies research group, which organises academic workshops[35] allied to major international conferences in the LT field, publishing papers in relevant peer-reviewed proceedings.

Supercomputing Wales[36] is the national supercomputing research facility for Wales. Its constituent members are Cardiff University, Swansea University, Bangor University as well as Aberystwyth University. Its programme includes investment in two upgraded supercomputer hubs and a new group of Research Software Engineers across Wales to develop algorithms and customised software that harnesses the power of the facilities. The facilities are supported by an experienced and specialised technical support team providing wrap-around services. To date these facilities have been used to train speech models for Welsh and it is anticipated that greater use of them will be made of them in the near future.

In line with increased LT research activity in Wales, concerted efforts have also been made to improve teaching and learning of digital technologies. In Wales, Information and Communication Technology is taught as part of the National Curriculum as a compulsory subject for Key Stage 2 pupils (those aged 7-11) and Key Stage 3 pupils (those aged 11-14) (Welsh Government, 2008). Moreover, digital competency, along with literacy and numeracy, is an important mandatory cross-curricular element within the Welsh National Curriculum as

[29] https://gov.wales/sites/default/files/publications/2019-02/prosperity-for-all-economic-action-plan.pdf
[30] http://techiaith.cymru
[31] http://termau.cymru
[32] http://corpws.cymru
[33] https://rhwydwaith.techiaith.cymru
[34] http://techiaith.cymru/books/?lang=en
[35] See for instance https://aclanthology.org/volumes/W14-46/
[36] https://www.supercomputing.wales

a whole.[37] In order to offer guidance and support to those delivering the Curriculum the Digital Competence Framework has been published by the Welsh Government. As a cross curricular framework, it 'is meant to sit alongside ICT and computer science and encourage the integration of digital skills across the full range of lessons' (Champion, 2016). Teaching of LT is also being expanded at university level, with a new taught Masters in Language Technology established at Bangor University in 2020, and further new courses in LT are currently in development.

Some of the most important projects/applications in the field in the last five years

After years of small-scale and fragmented initiatives, the publication of the Welsh Government's Welsh Language Technology Action Plan in 2018 provided a coherent and planned way forward for the development of Welsh LT resources, tools and services. A major new 'Technology and the Welsh Language' project began at Bangor University in 2020, annually funded by the Welsh Government, which addresses most of the headings of the Welsh Government's Action Plan. This project aims to produce a suite of new resources, tools and services for Welsh, to be published under open, permissive licences as far as possible. The philosophy behind the project is that some of the core components or LT 'building blocks' can be created in the spirit of the Basic Language Resource Kit (BLARK) (Krauwer, 2003), with special reference to the Welsh Language Technology Action Plan. These components are grouped together for NLP, Speech Technology, and Translation Technology. A few are public facing, including a free version of a popular spelling and grammar checker, a Welsh transcriber, and Welsh personal assistant, but most are aimed at the business sector, where it is hoped that they will be picked up, used and reused in various software products, both by local companies in Wales and by international companies dealing with multilingual applications. 21 of these outputs have already been included in the European Language Grid (ELG), with nine more promised by the end of March 2022. As noted in Section 4.2, one of the most important developments under this project has been the publication of a translation memory sharing platform, in emulation of the European Language Resource Infrastructure (ELRI) project.

A free version of the language tool compendium Cysgliad was also included in this project, as such resources are considered vital to confidence building in writing in a minoritised language environment. The spelling and grammar component, Cysill, had already been adapted as a free online tool, enabling the collection of a large (more than 200,000,000 tokens to date) corpus of contemporary written Welsh. Now it has also been adapted for use within Microsoft and Google online tool environments, showing the importance of updating existing language tools as technology moves forward. Cysill was written following older rule-based system, and is just one example of where newer neural net methodologies are needed to create a new generation of improved language tools.

In addition to producing the resources, tools and services themselves, the Technology and the Welsh Language Project also contains a large element of engagement activities with various stakeholders. The establishment of a National Network for LT in Wales was one of these, but there is also a programme of conferences, workshops and outreach events organised to disseminate project results and to encourage the take-up of resources and tools.

Another significant project during the last five years was the CorCenCC corpus project referred to in 4.1 above, funded by 2 UK Research Councils, the AHRC and ESRC. It has been used as the basis for a collection of data-driven teaching and learning tools (Y Tiwtiadur) designed to help supplement Welsh language learning at all different ages and levels, and, along with a selection of other corpora, to create other language resources such as a set of Welsh language word embeddings.

[37] https://www.gov.uk/government/news/new-ten-year-plan-to-make-britain-a-global-ai-superpower

As a minoritised language outside the mainstream of official EU languages, Welsh has not had much opportunity to take part in major international cross-language projects. Two notable exceptions however are Mozilla's Common Voice (CV) project[38] and the Welsh language Wikipedia. Common Voice aims to collect a sizable body of recorded speech under a CC-0 licence for use in developing speech technology for a large number of languages. Welsh was one of the three first additional languages, together with French and German, to be included in this project after the initial English version. As noted in Section 4.1, 143 hours of Welsh recordings have been made by over 1,600 volunteers,[39] by far the largest and most comprehensive such dataset for Welsh. This dataset has been used to create improved language models for Welsh, included in the Trawsgrifiwr speech transcription service. The models are updated and improved with every new CV release, showing the power of such international efforts to help small LT communities reach beyond what would have been possible with their resources alone.

As part of the Welsh LT Action Plan the Welsh Government have committed to support Welsh language Wikipedia editing workshops. In co-operation with the Welsh National Library and Menter Môn separately, Wiki workshops designed to help people edit and write Wikipedia articles through the medium of Welsh have been held. Moreover, Wikimedia has a dedicated UK Manager in Wales who has strongly promoted Welsh language resources for Wikimedia, Wikipedia and Wikidata, resulting in much improved language resources for Welsh, both in terms of quality and quantity (Welsh Government, 2020),[40] that are also reusable in other LT contexts.

LT providers who conduct research or offer tools/services for the language

The LT industry in the UK is heavily orientated towards serving the needs of the English language and other commercially lucrative markets. LT providers for Welsh in Wales exist mainly in the university sector, with Bangor University the main player, followed by Cardiff University,[41] and the University of South Wales' Hypermedia Research Group.

Wales does however have a vibrant creative technology and media sector, driven mainly by the commercial opportunities offered by S4C, the Welsh language television channel. This sector in turn nourishes a number of small, local companies producing software and apps for Welsh or bilingual markets. The provision of open-source resources and tools through government-funded projects is intended to encourage such companies to increase their activity in this area, thus also improving their economic viability. There is also a vibrant translation sector in Wales (Prys et al., 2009), partly in response to Welsh national bilingual policies and the need for bilingual documents for the public sector. This has in turn led to opportunities for others to service the need for translation tools, which are in common use throughout the industry. Cymen Cyf., a translation company based in Caernarfon, Gwynedd, has taken advantage of research funding to develop new translation technology, including MT systems trained on the company's own legacy translation memories (Prys and Jones, 2019). An emerging important player in the business landscape for Wales is MSparc[42] a Science Park on the Isle of Anglesey which has ambitions to promote LT and AI industries within Wales through the medium of Welsh or English. It works with local companies to support new ventures and through its Skills Academy to develop their technological expertise.

[38] https://commonvoice.mozilla.org
[39] Figures correct as of December 2021
[40] Note that the number of Welsh language Wikipedia articles had increased from 35,807 in 2012 to 131,002 by December 2020.
[41] Including Cardiff University's School of Welsh, School of Computer Science and Informatics, School of Mathematics, and School of English, Communication and Philosophy, in no particular order.
[42] http://www.m-sparc.com

Important tools and services for Welsh are also provided through crowdsourcing activities. Volunteer localisers have translated a range of open-source software such as LibreOffice, Firefox and Wordpress, into Welsh. The website Meddal.com (available in Welsh only) operates as a portal and provides comprehensive lists and links to localised products.

5 Cross-Language Comparison

The LT field[43] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded results unforeseeable before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

5.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[44] broadly categorised into a number of core LT application areas:
  - Text processing (e. g. part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g. search and information mining)
  - Translation technologies (e. g. machine translation, computer-aided translation)
  - Natural language generation (e. g. text summarisation, simplification)
  - Speech processing (e. g. speech synthesis, speech recognition)
  - Image/video processing (e. g. facial expression recognition)
  - Human-computer interaction (e. g. tools for conversational systems)
- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

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

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

5.2 Levels of Technology Support

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

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in ≥3% and <10% of the ELG resources of the same type
3. **Moderate support**: the language is present in ≥10% and <30% of the ELG resources of the same type
4. **Good support**: the language is present in ≥30% of the ELG resources of the same type[45]

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[46] and, on top of that, more than 6,000 additional language resources and tools were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the Digital Language Equality (DLE) metric will allow for dynamic analyses and calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[47]

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

[45] 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.

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

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

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

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

5.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 2 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,[48] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

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

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

[48] In addition to the languages listed in Table 2, 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.

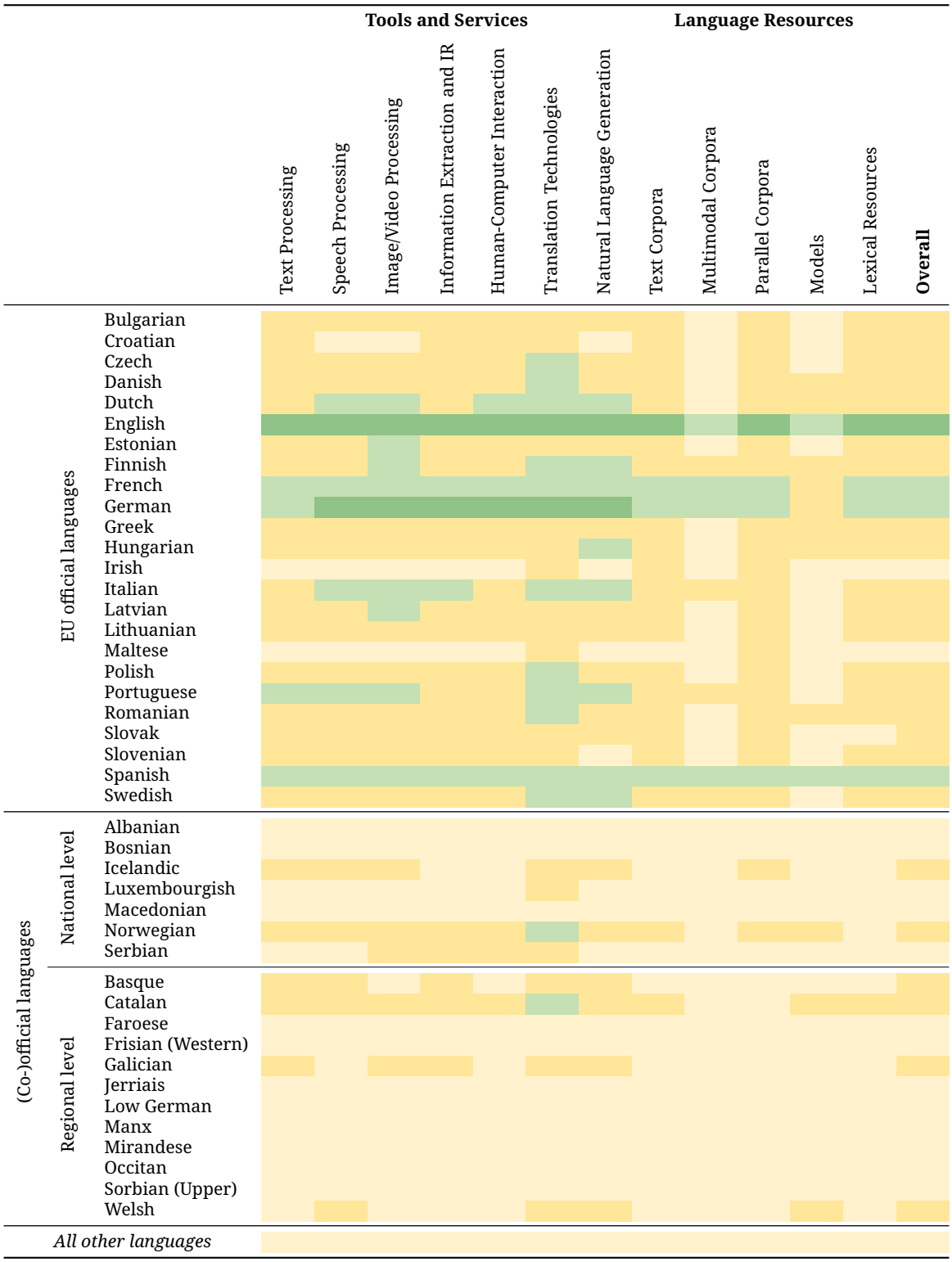

Table 2: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

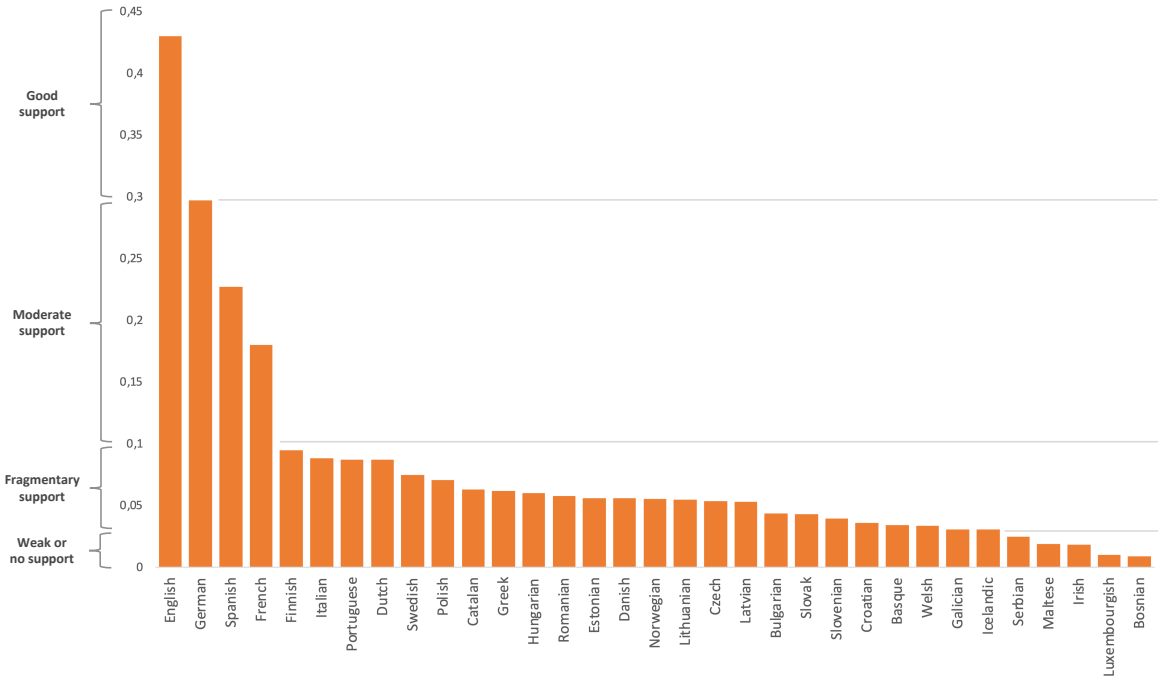

Figure 1: Overall state of technology support for selected European languages (2022)

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

Strengths and weaknesses

Although the tools, resources, services and products detailed in this document may seem impressive for a relatively small language with no official EU recognition, many types of resources, tools, services and products remain under-represented or need further development to attain parity with those of major languages. Long-term sustainability and further development of these resources and tools is problematic, in common with other minoritised languages. There is a culture of giving 'one off' grants to develop new tools and resources. However, less attention is paid to maintaining and updating them, and they can quickly become obsolete. Lack of commercial support also impedes progress, despite the potential for using LT development to grow local companies.

However, Welsh has made considerable progress in the digital sphere in the last few years, especially following the adoption of a Welsh Language Technology Plan for Welsh. The adoption of explicit official status for it in Wales, co-official with English, has also improved its prospects, and bilingual policies in official bodies have led to the availability of large-scale bilingual corpora which may be reused for developing MT and other technologies. Crowdsourcing has proved to be an effective way of creating speech corpora, and in common with

many other minoritised languages, Welsh has benefited from a coherent pool of language activists who are keen to contribute to building new resources needed for LT and AI purposes.

Progress has also been made in LT and digital education in Wales, from improving the National Curriculum in schools to introducing new courses at university level. Research programmes have been established, although funding is modest by international standards, and still dependent on short term budgets. There is a good understanding of the benefits of open, permissive licences for LT resources and tools, and the drive to disseminate widely, update and reuse in a frugal manner has helped a small language community make the most of scarce resources.

Efforts have been made to contribute to the development of the Welsh economy by encouraging the uptake of LT resources and tools by the software, creative industries and translation sectors. These efforts are gaining traction, but more remains to be done. Increasing understanding of the potential economic value of Welsh LT is work in progress, but mechanisms such as the National LT Network and biannual Technology and the Welsh language conferences are helping to improve communication between academia and industry. Integration into international LT and AI activities remains weak. As a minoritised language community with a small population base, Welsh is of no great commercial value to many multinational and multilingual companies.

Statutory requirements for bilingual tools and services in Wales redresses this balance somewhat, as does the provision of resources on permissive, open licences, thus bringing down the cost of including Welsh in multilingual tools and service provision. However, the Welsh Government and other public sector bodies heed the call of the Welsh Language Commissioner (Welsh Language Commissioner, 2021b) to use its purchasing power to insist on Welsh-language provision. Licensing these datasets for reuse in LT tools and resources is a key requirement for the advancement of Welsh language LT.

It is worth noting when the UK was a member of the EU, Welsh did not have an official status at the EU level, and thus was not able to participate in many of the projects reserved for the official languages, although it benefited from Interreg, ERDF and ESF funding. As such it might seem that Brexit would have no great impact on the prospects of LT for Welsh. However, as part of the European linguistic landscape, with its link to other Celtic languages and other minoritised languages, retaining links with those communities and to the European mainstream is of paramount importance. Welsh risks being left behind in LT as well as in cultural and other spheres if its bonds to the rest of Europe are not strengthened.

Gaps that a large-scale LT R&D programme could fill

Despite the considerable progress made in recent years, the Welsh language LT community has further work to do if the Welsh language is to thrive in this area. For instance, while FAQ generation is currently available and in use for the Welsh language the development of more sophisticated chatbot systems would further benefit Welsh speakers. Currently there is no published research on Welsh language knowledge graphs, nor what such technology could offer the Welsh language. Limited work and research has been conducted on Welsh language sentiment analysis. A key new area for development is bilingual models to aid minoritised languages such as Welsh where users constantly have to switch between their own language and the majority language (English in the case of Welsh). Promising work has been done for Welsh in developing a bilingual model for Text to Speech in the Lleisiwr project, enabling users to record their voices once for both languages, and similar work for speech recognition is underway. There are many other bilingual situations where a similar approach could be explored.

In order to fill the gaps for LT provision for Welsh, two seemingly contradictory paths need to be followed. Firstly, Welsh needs to be enabled to join in large-scale multinational

and multilingual R&D programmes of the type previously reserved for official EU languages, while secondly, and in common with other minoritised and other 'small' languages, its needs a space within the European community where special attention can be paid to up-resourcing these languages and up-skilling their language communities. Minoritised European languages often also belong to the economic periphery in Europe, and using LT for economic regeneration in those areas would have a positive effect on the economic as well as the social and linguistic well-being of those areas.

It is often more attractive to chase after new and exciting project ideas, and funding opportunities are often prejudiced in favour of such ventures, attention needs to be paid to improving, consolidating and further developing existing tools and resources. At the same time minoritised languages such as Welsh need to take full advantage of any emerging innovative solutions, playing their full part in the LT developments for Europe.

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.

Tegau Andrews and Gruffudd Prys. Terminology standardization in education and the construction of resources: The Welsh experience. *Education Sciences*, 6(1):2, 2016. ISSN 2227-7102.

BBC Cymru Wales and S4C and Welsh Government. Exploring Welsh speakers' language use in their daily lives. Report, BBC Cymru Wales, S4C and Welsh Government, 2013.

British-Irish Council. Report of the promoting our languages through technology conference. Report, British-Irish Council, 2017.

Klara Ceberio, Antton Gurrutxaga, Claudia Soria, Irene Russo, and Valeria Quochi. How to use the digital language vitality scale, 2018.

Joe Champion. Curriculum reform: The digital competence framework, 2016.

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

Daniel Cunliffe, Andreas Vlachidis, Daniel Williams, and Douglas Tudhope. Natural language processing for under-resourced languages: developing a Welsh natural language toolkit . *Computer Speech & Language*, 72, 2021.

Dewi Jones, Delyth Prys, Myfyr Prys, and Gruffudd Prys. Llawlyfr technolegau iaith, 2019.

D. Knight, S. Morris, T. Fitzpatrick, P. Rayson, I. Spasić, E-M. Thomas, A. Lovell, J. Morris, J. Evas, M. Stonelake, L. Arman, J. Davies, I. Ezeani, S. Neale, J. Needs, S. Piao, M. Rees, G. Watkins, L. Williams, V. Muralidaran, B. Tovey-Walsh, L. Anthony, T. Cobb, M. Deuchar, K. Donnelly, M. McCarthy, and K. Scannell. Corcencc: Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh, 2020.

Steven Krauwer. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *SPECOM 2003*, pages 8–15, 2003.

Ceri W Lewis. *Orgraff yr Iaith Gymraeg: Geirfa Rhan II*. University of Wales Press, Cardiff, 2018.

Daithí Mac Síthigh. Official status of languages in the UK and Ireland. *Common Law World Review*, 47: 25, 2018.

Christopher Moseley. *Atlas of the World's Languages in Danger*. UNESCO Publishing, Paris, 2010.

National Assembly for Wales. Welsh language (Wales) measure 2011, 2011.

Ofcom. Connected nations 2020 Wales report. Report, Ofcom, 2020.

Delyth Prys. *Setting the Standards: Ten Years of Welsh Terminology Work*, pages 41–55. Gunter Narr, Tübingen, 2006.

Delyth Prys, Gruffudd Prys, and Dewi Jones. Improved translation tools for the translation industry in Wales : an investigation final report. Report, Bangor University, 2009.

Gruffudd Prys and Gareth Watkins. *Welsh Word2vec model: vector representation of the semantic correlation of Welsh words based on their embeddings within an enormous Welsh corpus*, volume 1 of *Language and Technology in Wales*, book section 8, pages 87–107. Bangor University, Bangor, Wales, 2021. ISBN 978-1-84220-188-6.

Myfyr Prys. *Style in the vernacular and on the radio: code-switching and mutation as stylistic and social markers in Welsh*. Thesis, 2016.

Myfyr Prys. *Implementing NMT at a Welsh translation company*, volume 1 of *Language and Technology in Wales*, book section 9, pages 107–120. Bangor University, Bangor, Wales, 2021.

Myfyr Prys and Dewi Bryn Jones. Embedding English to Welsh mt in a private company. In *Celtic Language Technology Workshop. European Association for Machine Translation*, pages 41–47, 2019.

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.

Jiao Song, Catherine A. Sharp, and Alisha R. Davies. Population health in a digital age: Patterns in the use of social media in Wales. Report, Public Health Wales & Bangor University, 2020.

StatsWales. Annual population survey - ability to speak Welsh by local authority and year, 2021.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423.

University of Wales. *Orgraff yr iaith Gymraeg : adroddiad*. Gwasg Prifysgol Cymru, Cardiff, 1928.

Welsh Government. Information and communication technology. Report, Welsh Government, 2008.

Welsh Government. Welsh language use in Wales, 2013-15 . Report, Welsh Government, 2015.

Welsh Government. Cymraeg 2050 a million Welsh speakers, 2017a.

Welsh Government. Welsh second language in the national curriculum for Wales, 2017b.

Welsh Government. Welsh language technology action plan, 2018.

Welsh Government. National survey for Wales: Headline results, april 2018 – march 2019. Report, Welsh Government, 2019.

Welsh Government. Welsh language technology action plan: Progress report 2020, 2020.

Welsh Government. Internet skills and online public sector services (National Survey for Wales): April 2019 to march 2020. Report, Welsh Government, 2021.

Welsh Language Commissioner. The Welsh Language Measure, 2021a.

Welsh Language Commissioner. The Position of the Welsh Language 2016–20: Welsh Language Commissioner's 5-year Report. Report, Welsh Language Commissioner, 2021b.