CYMRAEG

Llywodraeth Cymru
Welsh Government

Welsh language technology action plan

Welsh language technology action plan

Audience

All those interested in ensuring that the Welsh language thrives digitally.

Overview

This Welsh language technology action plan derives from the Welsh Government's strategy *Cymraeg 2050: A million Welsh speakers* (2017). Its aim is 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.

Action required

For information.

Further information

Enquiries about this document should be directed to:
Welsh Language Division
Welsh Government
Cathays Park
Cardiff
CF10 3NQ

e-mail: cymraeg@gov.wales

@cymraeg

Facebook/Cymraeg

Additional copies

This document can be accessed from gov.wales

Related documents

Prosperity for All: the national strategy (2017); *Education in Wales: Our national mission, Action plan 2017–21* (2017); *Cymraeg 2050: A million Welsh speakers* (2017); *Cymraeg 2050: A million Welsh speakers, Work programme 2017–21* (2017); *Welsh-language Technology and Digital Media Action Plan* (2013); *Technology, Websites and Software: Welsh Language Considerations* (Welsh Language Commissioner, 2016)

Mae'r ddogfen yma hefyd ar gael yn Gymraeg.
This document is also available in Welsh.

 WG34015 Digital ISBN 978 1 78964 010 6

Contents

1. Ministerial foreword

I use technology every day. It's central to all our lives, whether that technology is visible to us or not. Technology is everywhere, but I am seldom offered the opportunity to use it in Welsh. And if a Welsh Language option is available, it is not always visible without me making an effort to look for it. We all lead busy lives—who has time to go looking?

Technology has already transformed the way we live, and it's quite possible that change will be more rapid as time goes on. *Cymraeg 2050: A Million Welsh Speakers*, the Welsh Government's strategy for the Welsh language, recognises that such transformations create challenges for the Welsh language. However, they also bring us opportunities. Indeed this is the case for smaller languages the world over.

I strongly believe that we need to grasp these opportunities and tackle technological challenges—and make sure that Welsh is a central part of the agenda. That's what I intend to do by publishing this *Action Plan* for Technology and the Welsh language. The plan underpins *Cymraeg 2050* and intends to foresee wider technological developments, and so set the direction of travel for work on Welsh language technology.

But technology for technology's sake is of no interest to me; as omnipresent as technology may be—that's not what's important. Rather, if technology can bring more opportunities to live in Welsh, or to learn Welsh, we need to grasp them. If technology can make it easier to work and provide services in Welsh, we need to develop opportunities to do those things. If technology can improve quality of life for those living with accessibility challenges, we need to develop that technology in Welsh. And all this provision needs to be widely used.

I'd like to see a situation where Welsh is used in as many situations as possible—education, work and home, and in all those situations which technology reaches. I already use voice technology on my phone. I'd like to be able to do that in Welsh. Like me, most Welsh speakers live or work in a bilingual atmosphere. So as I work, the technology I use will have to be able to handle Welsh and English at the same time.

This is an ambitious document, and it will be implemented by more than the Welsh Government alone. But we won't shy away from leadership or financing developments where that's what's needed. I'm grateful to the Board of specialists which I chair for all their input and to all those who provided advice during the development of this Plan. We will only realise this Plan's aims by collaboration, and thereby lay one of the essential foundations for a million speakers by 2050.

I’m conscious that technology develops quickly. I’m eager for the Welsh language to move with it. That is the aim of this Plan.

ELUNED MORGAN AM
Minister for Welsh Language
and Lifelong Learning

2. Vision

Cymraeg 2050 notes the importance of technology for the future of the Welsh language, and the necessity for technology to support Welsh, for the language to be used in as many situations as possible.

Very few of us, particularly children and young people, do not have a daily connection with technology. This *Action Plan* recognises technology as a priority area to be addressed in order to ensure a place for Welsh in our future lives. Technology is a game changer in 2018.

This Plan does not start from scratch; it builds on foundations already laid—by volunteer technologists, by large enterprises and by the public sector itself. What the Plan does do is assemble the main language technology components and projects in order to create a national ambition for language technology. This ambition will support current and future Welsh speakers—in education, and in their professional and social lives.

We have identified three specific areas to be addressed by the plan:

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

The plan clearly sets out the challenges for these areas. It pays attention to infrastructure, and use, but our vision brings components together so that language technology materials may be available to use and to share for the future.

To summarise: we want to see the Welsh language at the heart of innovation in digital technology. We want to be in a situation where Welsh is offered proactively, and in which families, organisations and individuals can use Welsh in an increasing number of digital contexts—be they monolingual or multilingual—without having to request it.

To realise this vision, we have to take action to ensure sustainable provision for future generations. This long-term process starts with this *Action Plan*, and we set out principles for achieving this below. A glossary defining the terms used is included at the end of the document. The main work required by this Plan will be undertaken in the following fields:

We will do this by:

Creating and sustaining Digital Infrastructure
We must ensure that we have the necessary Welsh-language digital infrastructure we need, such as speech-to-text, machine translation, relevant corpora and big datasets, machine learning etc. We must have Welsh-language data to train these systems. We need to make sure that academic research and development in the field of Welsh language technology receives long-term support. We need to maintain and update this infrastructure for future generations.

Developing a Culture of open Innovation
We want to see more Welsh-language digital resources and data openly available, without unnecessary conditions limiting their use. We want to see those resources shared under an appropriate licence—that is open and permissive whenever possible—to support the digital innovation effort that benefits the Welsh language. This will have an effect on education, and the resources shared will encourage people to create and publish Welsh-language content. Where possible, we will favour generic international standards with regard to everything we do in the context of this *Action Plan*.

Our aim is that this innovation and development and the emphasis on open technology and data will contribute to an Innovative Economy in Wales. This plan provides an opportunity for Wales to develop the technology sector by using our experience and language technology skills to work with other nations and regions who also want to see speech-to-text, improved machine translation, conversational AI, etc. in their own languages.

Building Capacity and Digital Skills
We need more resources and learning opportunities for people of all ages to facilitate digital work and coding in Welsh as well as learning how to handle Welsh-language technologies and data. We need to create the language technologists and computational linguists of the future to suit the kind of work that will be available in the future.

Digital Transformation in the Public Sector
The Welsh language has to be an integral part of the digital transformation agenda in the public sector. Through the power of procurement and other linked processes our aim is to influence the wider ICT sector to offer more Welsh language provision.

Promoting the creation and use of Welsh Language Digital Products and Services
Welsh-language interfaces, resources, and products must be available in a friction-free manner, without having to be requested by the end user. Choice architecture methodologies should be used to expedite this. Considering the number of bilingual English/Welsh-speaking households and workplaces in Wales at present, where Welsh speakers live or work with non-Welsh speakers, Welsh language provision must work smoothly in bilingual households, workplaces and all types of situations.

Welsh-language user experiences must facilitate the finding of Welsh-language content. When introducing resources and content, our aim would be to ensure a wider range of Welsh-language digital products and services, and to increase the use of those resources.

3. Summary

This section provides an overview of those fields which this *Action Plan* addresses. A glossary defining the terms used is included at the end of the document.

Over the lifetime of this plan we will, with relevant partners, work towards developing the following:

- Welsh language speech technology.
- Conversational artificial intelligence (conversational AI) for the Welsh language.
- Improved user experience (UX) for users of Welsh-language technology.
- Encouragement for technology companies to use more Welsh.
- Voice banking facility for individuals' Welsh language voices.
- Relevant technical learning resources for those learning Welsh.

In our Education and Skills work, we will:

- Take advantage of the new curriculum and the Hwb website[1] to develop children's and young people's skills in digital literacy, coding, digital content creation, etc. in Welsh.
- Examine the potential of developments such as e-sgol for increasing Welsh-medium educational provision
- Promote wide and proactive use of Welsh-language interfaces and software on devices for Welsh-speaking learners and staff at Welsh-medium schools, colleges and universities in Wales.
- Promote of Welsh-language coding and other relevant resources.

For workplaces and services, we will:

- Examine the potential of developing automatic translation systems.
- Facilitate more sharing of translation memories.
- Adapt relevant technology procurement procedures to ensure the Welsh language is considered from the outset.
- Facilitate the proactive offer of Welsh language choice in services and interfaces.

To support Welsh language content creation, we will:

- Support workshops that ensure people create and publish video, audio, image and text content in Welsh.
- Support terminology, lexicography and corpora resources and other elements of Welsh language infrastructure.
- Help to improve Welsh spellcheck, grammar and mutation checkers and aim to make more such facilities available free of charge.

[1] http://hwb.gov.wales

4. Introduction

The Welsh language is not the only language facing the need to develop a language technology infrastructure. For example, in 2013, META-NET foresaw that at least 21 European languages were in danger of digital extinction.[2] Even though META-NET noted Welsh[3] as one of these languages, it merits stating that there are 'larger' languages, with a far higher number of speakers, that are also facing similar challenges to those outlined below, namely: finding a digital foothold in a world where the English language is becoming ever more prominent.

One of the challenges we face in Welsh language technology is that 59% of the 550,000 people who are Welsh speakers live in households with at least one non-Welsh-speaking person (according to the 2011 Census).[4]

As a result, not only do we have to ensure that appropriate Welsh-language technology exists, we have to work towards ensuring that Welsh and English technology work simultaneously on the same platforms. For example, it would not be ideal to expect people to buy separate Welsh and English home smart digital assistants. Ensuring that both languages work on the same popular products means discussing with large companies with a view to ensuring that the existing and new Welsh language components are easily accessible on their systems. Therefore, when we implement this plan, we will set our sights on doing more than merely creating a language technology infrastructure (such as Welsh-language speech-to-text technology and improved machine translation). Developing relationships with developers and organisations that create the technologies of the future should help ensure that the Welsh language is available in as friction-free a manner as possible for every individual user.

When we discuss educational resources, for example, digital resources are an integral part of this. The same applies for resources to help Welsh speakers in the workplace, and when considering services and products. Of course, it is important to emphasise that this plan has a central role to play when it comes to driving forward the accessibility through technology agenda for Welsh speakers facing accessibility challenges.

Having a presence on the internet raises the profile of the Welsh language. Ventures that enable the creation of Welsh-language digital content can contribute to this by helping school pupils to create and surface Welsh-language content, and use it outside school. To that end, the Digital Competence Framework[5] will continue to offer tools and resources in Welsh and develop ways to facilitate creation of more

[2] META-NET (no date), White Papers Series, http://www.meta-net.eu/whitepapers/overview, META-NET, Berlin.
[3] META-NET, (2013), *The Welsh Language in the Digital Age,* META-NET White Papers Series, META-NET, Berlin. http://www.meta-net.eu/whitepapers/volumes/welsh.
[4] Welsh Government (2013), *Welsh Language households and transmission*, Welsh Government, Cardiff, https://llyw.cymru/docs/statistics/2013/130628-welsh-language-households-transmission-en.pdf.
[5] Welsh Government, 2018, *Digital Competence Framework* http://learning.gov.wales/resources/browse-all/digital-competence-framework/?skip=1&lang=en

Welsh-language content and the use of every kind of technology through the medium of Welsh.

The current document was compiled jointly by the Welsh Government, following input from various sources. Firstly, the Welsh Government's and the Welsh Language Board's previous digital strategies for language technology were reviewed. Desk research was also undertaken on other relevant documentation.[6] The Minister's Welsh Language Technology Board provided advice in workshops (Board Members are listed in Appendix 3). The resultant draft priorities were sent to the members for comment as was the draft plan itself. Members of the Hacio'r Iaith community also readily provided input.

In terms of the implementation of this *Action Plan*, we will continue the 'co-creative' approach adopted for its development. That means collaboration and partnerships will be a key way of working. However, wherever the Welsh Government is the lead organisation for a relevant work package, methods available to us may include developments via procurement or grants. We will also use promotional, marketing, and any other relevant avenues.

[6] For example, see The Digital Language Diversity Project (2018), *Digital Language Survival Kit: The DLDP Recommendations to Improve Digital Vitality*, DLDP, http://wp.dldp.eu/wp-content/uploads/2018/09/Digital-Language-Survival-Kit.pdf

5. Work packages

Below, we set out a list of work packages that are critical to realising the vision of this *Action Plan*. It is set out in a consistent format, stating the area of work, the reason for its inclusion, how we hope to achieve the work and the risk of not taking that action. Our aim is to be as concise as possible. This section should be read in conjunction with the terminology Glossary at the end of this document.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
1	Welsh-language speech-to-text facilities and components.	Up until comparatively recently, the keyboard has been the main input method into computers; today voice is more important than ever. Ensuring Welsh language speech-to-text facilities will mean we can issue verbal commands and obtain automatic transcriptions of what we say and of, for example, sound and video archives.	Ensuring a speech to text engine of sufficient quality and training it with relevant data sets and corpora. Where possible, complement crowdsourcing work already undertaken in this field (e.g. Mozilla Common Voice). A continuous, long-term programme of research and development will be needed in this rapidly-changing arena with the advent of neural networks.	Software created under an appropriate licence, work undertaken to convince relevant organisations/corporations to adopt that software.	Without further developments in this area, it will not be possible to ask spoken questions of Welsh to computers, machines or phones. There will be no means of managing Welsh-language information and technology systems through voice (all the while, an increasing number of English language systems are adopting voice technology). Also, by not developing Welsh-language speech recognition technology, there will be accessibility and availability implications for those Welsh speakers with additional needs.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
2	Welsh-language machine learning and conversational Artificial Intelligence (AI).	AI can drive robots and machines to build things e.g. cars, but this work package deals with natural language understanding for Welsh. This will enable computers to analyse Welsh language questions and to respond in a sensible and useful way. Conversational bots and other technologies depend on this infrastructure. Such bots can be used, by citizens, for example, to make enquiries of their local council, or to book a course at night school (amongst a myriad of other possibilities).	Examine frequently asked questions; in order to feed these into a relevant bot for the Welsh language. Plan a programme of analysis and extraction of information from suitable sources.	Software under an appropriate license.	Individuals and families would have to continue to use English to ask questions of machines. Before we can create conversational AI and speech facilities, we need to have lexicons, corpora of parallel texts, stemmers, parsers, term extraction, named entities, sentiment analysis scores, lists of stopwords and WordNets. More details on these can be found in a technical Appendix below. To train systems, we will need transcribed audio corpora.
3	Welsh language audio corpora for annotation, to train speech-to-text and other technologies.	To train speech-to-text systems and to work towards having automatic transcription in Welsh (and as a result have the ability to translate and write subtitles simultaneously) with the help of computers.	These audio corpora and other data sets to be created with the co-operation of broadcasters, archives, public organisations and other relevant bodies.	A collection of transcribed audio recordings suitable to train systems.	There may be a lack of accuracy and maturity with regard to Welsh speech-to text engines.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
4	Improving user experience of Welsh language technology through behavioural economic techniques.	Evidence shows that the number of users using Welsh language technologies and computer services is relatively low. Research[7] shows that adopting behavioural economic techniques (and other methods) may increase these numbers.	Our focus will be on how to create a friction-free Welsh language choice architecture (see also below in the context of tracking and recording language choice). We will use a mixture of methods to do this, which may include: • User experience techniques, A/B testing, focus groups, surveys etc. • We will also ensure that good practice guides are disseminated with regard to optimum language choice architecture techniques which apply the best principles of human-computer interaction, behavioural economics and other relevant research to increase the use of Welsh. • This guidance will be at the heart of the Welsh Government's *Understanding Bilingualism* programme[8] and will be co-created with relevant partners from public services, industry and others.	Research Reports, practical guides, techniques to increase the number of the Welsh-language *locale* users (be they individual users or institutional users), indicators of users that use Welsh services/interfaces.	If the steps outlined above are not taken, the number of people using Welsh-language services may not increase.

[7] See, amongst others, Evas, J.; Cunliffe, D. (2016), "Behavioural Economics and Minority Language e-Services—The Case of Welsh", in *Sociolinguistics in Wales*, eds. Morris, J. and Durham, M., Palgrave MacMillan, London.
[8] This is a long-term programme of critical Welsh language awareness and other related projects.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
			• This, and our wider work, will compare the effect (with regard to increasing the number of Welsh language technology users) of using a Welsh language *locale* on devices rather than requiring users to change individual programme settings. Our goal, as with every aspect of our technological work, is to ensure a friction-free Welsh language computer experience for the user.		
5	Frameworks to personalise text-to-speech and bank individual voices.	So that Welsh speakers who are in danger of losing their voices for health reasons will still be able to communicate orally in Welsh, with their own personalised speech synthesis engine.	Assess how to develop technology such as Bangor University's pilot synthetic voice banking project funded by the Welsh Government.	Software and instructions available under appropriate licence.	If such technology is not available, Welsh speakers who have lost their voice may only be able to communicate orally via English language technology or by using a non-personalised synthetic voice.
6	Interactive content and software for Welsh learners.	One of the main strategic aims of the Welsh Government Strategy, *Cymraeg 2050* is to increase the number of Welsh speakers. Welsh learners have a key part to play in relation to achieving the Strategy's aims.	Work with the National Centre for Learning Welsh and other partners to pilot learning with new technologies such as chatbots, games, Virtual Reality and distance learning.	New digital resources available to help people learn Welsh and to improve oral and written Welsh.	a lack of suitable technological opportunities to enrich the experience of Welsh learners.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
7	**Education and Skills** The Welsh language to be the user interface (UI) language of devices in Welsh-medium education and for Welsh-speaking students and staff in colleges and universities in Wales.	So that students, pupils and staff can use a Welsh language UI and facilities to help create Welsh content automatically and in a friction-free manner when learning and working.	Guidelines and collaboration with networks representing relevant educational institutions, for example to link the linguistic records of student and personnel management systems to user ID management systems. 'Group policies' can thereafter be created to automate a Welsh UI for Welsh speakers, giving an option to opt out if they wish (as is generally done currently in the case of the default English language experience). Work may need to be done to ensure that toggling between English and Welsh can work without rebooting devices.	Lessons learned logs, case studies, a report on the Welsh language UI in a sample of organisations.	Missed opportunities to introduce Welsh language UI to Welsh-speaking members of educational institutions to increase the number of those who can have a Welsh language UI.
8	Promote Welsh language technology and coding resources to teachers and children and others.	To ensure that pupils, teachers, students and learners in Wales have the best skills for jobs now and in the future. Build capacity in Wales to use and further develop Welsh language technology and coding resources, and as a result, contribute to the processes of creating the computational linguists and language technologists of the future.	Work with Cracking the Code to market the resources already available and ensure there are no gaps in the Welsh language resources that are available through commissioning. Work with Cracking the Code, education consortia, Codeclub UK and others to support the localisation of resources used to learn coding. Examine the possibility of dedicated training for computational linguists. Undertake outreach work with relevant organisations/individuals to provide training on how to use Welsh language technology.	Relevant resources available, training carried out.	A lack of succession planning for computational linguists, a lack of Welsh language coding technology skills and a general lack of technology skills, lack of uptake of Welsh language interfaces.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
9	Create and/or develop facilities to assist Welsh speakers with additional learning and/or accessibility requirements.	To facilitate the use of the Welsh language in every possible circumstance and to remove barriers for people who have accessibility requirements.	List the relevant components available in English but not in Welsh and then examine how to fill these gaps.	Relevant resources available.	Welsh speakers with accessibility requirements may be at a disadvantage when learning and accessing services.
10	**The Workplace** Ensuring suitable English/Welsh and Welsh/English machine translation systems for different linguistic domains and registers.	A substantial body of research[9] has been carried into computer-assisted translation for Welsh and other languages.[10] That research is too extensive to summarize in this document, but one of its central themes is that human translators have been, and will be needed, to ensure translation quality, and the research	Commission, where appropriate, and/or work with relevant organisations to improve systems that already exist and disseminate their use.	Relevant training data available under an appropriate licence; Application Programming Interface to Welsh/English and English/Welsh translation engines available.	Were translation automation facilities for the use of human translators not to be developed, the Welsh language may not benefit from extant translation technology used for other languages and Welsh would not be as prevalent as it could

[9] http://techiaith.bangor.ac.uk/publications/?lang=en

[10] There are countless research publications available that analyse this area. Below, there is a selection of research papers on the issue of automated translation technology in the Welsh language context (they include literary reviews of relevant research done in the context of other languages). This list is not exhaustive:

Prys, D. 2014. *Advice Note*. Bangor: Bangor University Available from: http://techiaith.bangor.ac.uk/advice-note/?lang=en

Prys, D. *et al.* 2009. *Gwell Offer Technoleg Cyfieithu ar gyfer y Diwydiant Cyfieithu yng Nghymru: Arolwg Dadansoddol*. Bangor: Bangor University Available from: https://goo.gl/52ZYfj.

Screen, B. 2016. What does translation memory do to translation? The effect of translation memory output on specific aspects of the translation process. *Translation and Interpreting* 8(1), pp. 1-18. (10.12807/ti.108201.2016.a01)

Screen, B. 2017. Effaith Defnyddio Cofion Cyfieithu ar y Broses Cyfieithu: Ymdrech a Chynhyrchiant Cyfieithwyr Proffesiynol Cymraeg. *Gwerddon* 23(1), pp. 10-35.

Screen, B. 2017. Productivity and quality when editing machine translation and translation memory outputs: an empirical analysis of English to Welsh translation. *Studia Celtica Posnaniensia* 2(1), pp. 113-136. (10.1515/scp-2017-0007)

Screen, B. 2018. *Defnyddio Cyfieithu Awtomatig a Chof Cyfieithu wrth gyfieithu o'r Saesneg i'r Gymraeg: Astudiaeth ystadegol o ymdrech, cynhyrchedd ac ansawdd gan ddefnyddio data Cofnodwyr Trawiadau Bysell a Thracio Llygaid.* PhD Thesis PhD, Cardiff University

N°	What	Why	How	Deliverable	Risks of not undertaking this work
		evidence suggests that this is unlikely to change. The research also proves, on the basis of detailed evidence, that technology can increase a translator's productivity, assist in deleting repetitive translation, maintain consistency and speed up the translation process, whilst reducing the cognitive strain on the translator—all the while sharing translations between different organisations and translators in real time. Cymdeithas Cyfieithwyr Cymru (the Association of Welsh Translators and Interpreters) has also stated that it wants to make "the best and most effective use of technology" [11] in the field of translation. For many years, automatic translation facilities such as Google Translate and Microsoft Translator (amongst others) have been available. They are used to provide an automatic gist translation, and to this end, they assist non-			be in the linguistic landscape.

[11] Association of Welsh Translators and Interpreters, "Reforming local government offers an opportunity to establish strong translation and interpreting units", https://www.cyfieithwyr.cymru/en/newyddion (28/6/18).

N°	What	Why	How	Deliverable	Risks of not undertaking this work
		Welsh speakers in understanding Welsh language text. They are also being used to suggest sentences where no match exists in a human translator's translation memory system (see below). To assist human translators to increase the amount of Welsh-language material in Wales' linguistic landscape, we want to see further development of automatic translation systems.			
11	Take full advantage of existing translation memory software to assist human translators to increase the amount of Welsh-language material in the linguistic landscape. By using translation memories alongside appropriate machine translation it will be possible to share translations in real time.	To increase the amount of Welsh in the linguistic landscape in Wales.	Collaboration to ensure the use and sharing of existing translation memory systems and by ensuring new systems where appropriate and practical.	Shared computer-assisted translation networks; data on the use of technology and the translations thereby undertaken.	If action is not taken to create automation facilities for use by human translators, the Welsh language will not benefit from the extant translation technology used for other languages and Welsh will not be as prominent in the linguistic landscape.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
12	Modify, where relevant, procurement processes, so the Welsh language is a consideration in technology from the outset.	The Welsh language must be an integral part of the digital transformation agenda in the public sector. By using procurement processes, where relevant, we aim to normalise the ability to record the linguistic capabilities of systems procured. Where a given system cannot currently offer adequate Welsh language provision, our procurement process will help to drive the market to ensure change and influence on the wider ICT sector.	Relevant amendments to procurement systems so that a list of guidelines is available to ensure public sector technology suppliers know what they need to do to offer a Welsh-English bilingual service to citizens in Wales. Offer outreach sessions to explain the implications of the procurement system and the basics of a good bilingual computer system to businesses and industry.	Details on the Welsh Government's website that explains the amended procurement processes with attendant guidance.	Were this work not undertaken, all relevant apparatus would not be used to drive the public sector development of Welsh-language software by the technology sector.
13	Explore the potential of technology to facilitate and/or automate Welsh-language services e.g. automatically redirect phone calls to Welsh speakers within organisations.	So that Welsh speakers can have a service in their chosen language in as friction-free a way as possible.	Explore the potential of sharing language choice amongst service providers and using personnel and computer systems to route calls, messages and cases automatically to Welsh speakers.	White Paper reports and details on the number of services automating the choice of language and/or adopting technology that will enable them to do so.	If this is not done, we may miss strategic opportunities to develop technology that will offer more friction-free Welsh language services.
14	A list of Welsh-language and bilingual ICT resources available in the workplace, also noting gaps in provision.	So that Welsh speakers can make increasing use of Welsh language in the workplace.	Examine systems currently in use and note which support Welsh; conduct a gap analysis of those which do not in order to assess which Welsh-language technology services, programmes and computer resources are needed in workplaces.	A list of available and missing resources.	If this is not done, we will not possess as full an understanding as possible of what can be done to facilitate the use of Welsh in the workplace.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
15	**Welsh Language Content Creation** Support Welsh language Wikipedia editing workshops, video workshops and other channels that encourage people to create and publish Welsh-language video, audio, graphic and text content	To increase the presence of Welsh on the internet, to support the development of Welsh language Wikidata, to support Welsh-language activities which increase social capital and to encourage an increase of written Welsh. Videos on platforms such as YouTube are important to children and young people. We therefore need to examine what will work to target the content which is most needed in Welsh.	Work with a community of volunteers and support them.	Reports on the number of workshops and the resultant outputs.	That the Welsh language will not be used to its full potential in content creation if capacity building work to enhance Welsh language written and media production skills is not undertaken.
16	Long-term support for the development of the linguistic infrastructure of the Welsh language, including corpora, lexicographical and terminological resources.	These are the tools used worldwide to record, maintain and develop language infrastructure. These various resources dovetail to record a given language's history, usage and different forms, etc. They are used when developing resources for learners, language technologists, professional translators, in education etc. We will ensure that appropriate terminological and lexicographical sources will also be openly available under an appropriate licence wherever possible. This will facilitate consistency throughout the translation	Maintain and create relevant lexicographical/terminological resources. Regularly assess our activities in this area.	Downloadable resources released under an appropriate licence.	There may be a lack of suitable infrastructural resources available to implement the *Cymraeg 2050* strategy. Neglecting lexicographical and terminological work could lead to there being fewer approved terms and inconsistency in the Welsh used in different situations.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
		profession and in the wider Wales.			
17	Welsh spellcheckers, grammar and mutation checkers available free of charge.	Increase functionality of Welsh-language spellchecker/grammar checkers, ensuring they are freely available.	Undertake an options appraisal of grammar checkers/spellcheckers.	Relevant software available under an appropriate license.	If such resources are not made available, we cannot ensure that proofreading tools will reach the widest possible range of audiences possible to assist content creation in Welsh. Resultant opportunities to tackle lack of confidence to write in Welsh will be missed.
18	Interactive maps with Welsh language versions of place names that can be embedded within web pages.	So standardised Welsh-language place names can be used on maps and reach as many people as possible.	Create and develop a new layer of Welsh names on Open Street Map/other relevant platforms.	Relevant cartographic resources with Welsh language place names.	Welsh-language place names will not be viewed by users of digital maps and they may therefore not be customarily used.

6. Glossary

Term	Definition
Active offer (also *proactive offer*)	In the case of language policy, this is the process of offering a service in a given language without someone having to request that language.
Aligning/Alignment	The process of arranging text in corresponding parallel segments in Welsh and English. Audio recording can also be aligned with written transcript. Parallel, aligned content can be used to train translation memory, machine translation and speech-to-text systems.
API	Application Programming Interface: a means of enabling communication and data interchange between computer systems.
Chatbot/conversational bot	A 'Chat' engine which possesses conversational AI capacity.
Choice Architecture	How a particular choice is designed to be presented to end users, language choice being the focus of the current *Action Plan.*
Coding	Writing the code that is used to create computer software. This *Action Plan* addresses coding skills to build Welsh-medium capacity in this area.
Conversational AI	Some robots build items. Others use natural language processing to offer appropriate and useful responses to questions and circumstances, also known as AI (artificial intelligence).
Corpus (plural: corpora)	A large collection of texts, recorded or printed. Can also be a collection of sound recordings or human gestures (i.e. sign language).
Digital Competence Framework	Digital competence is one of three cross-curricular responsibilities for schools in Wales, alongside literacy and numeracy.
Digital literacy	A collection of different skills that are used to access, analyse and create digital products in various circumstances.
Domain	Subject, activity or information area. Examples of different domains are the law, science, or art. The ability to classify texts according to specific domains is important in many areas of Natural Language Processing, including Machine Translation.
Embeddings (word and term embeddings)	Machine learning algorithms that use corpora to assess a word within the context of a sentence and suggest a meaning based on probability.
Friction	A concept from Behavioural Economics. In the context of Welsh language services, an added 'cost' to the end user of using a Welsh language service, normally in time, frustration or cognitive strain. This *Action Plan* espouses friction-free Welsh language services *actively offered.*

Term	Definition
Gist Translation/Gisting	A machine translation which has not been post edited by a person. It gives the gist of the meaning, but is not intended for publication.
Internet of Things (IoT)	A network of devices, vehicles, buildings, and other items which include electronics, software, sensors, and network connectivity that allows these things to collect and exchange data.
Language Technology Infrastructure	The linguistic building blocks or components that enable us to build technological systems to process natural language.
Learning bot	See 'Chatbot/conversational bot'
Lemmatizer[12]	Morphologically analyses text and allows users to see the 'lemma' of any word that has been mutated, conjugated or inflected (see also 'stemmer'). For example, a lemmatizer knows that 'slept' comes from sleep.
Machine Translation	A service, e.g. Microsoft Translate/Google Translate, which can automatically translate text from one language into another. Such systems can be coupled with translation memory software. Machine translation is another example of a CAT (computer-assisted translation) tool.
Named Entities	These are useful to protect words and terminology and to ensure they are treated as a single entity by computer systems. People's names and place names are common examples of named entities. Lists of named entities are important in information extraction, from corpora for example. In machine translation, named entities are used to 'protect' units of meaning from being treated separately e.g. so that a person's name such as 'Dr Smith' is not translated literally into Welsh as 'blacksmith' ('Dr Gof').
Neural Networks	A way of processing data which mimics the human brain. Connections between artificial 'neurones' made and adapted during the training process.
Open and Open Licence	In the context of this document, open licences (there are several types of such licences) are referred to. These are used where the owner of the data or resource allows wide access, use and sharing. Permissive licenses can make it easier for companies to use the product without laborious terms and conditions.
Parser	Parsers 'reveal' the meaning of sentence semantics for machine translation, conversational AI etc. The computer analyses and splits a sentence and thus creates a type of 'tree', which reveals the grammar of the sentence, citing the conceptual relation of words to each other e.g. Dependency Parsers, Constituency Parsers

[12] For more Information, see: http://techiaith.cymru/api/lemmateiddiwr/

Term	Definition
Permissive License	Software license which confers the right to redistribute, alter and create proprietary derivative works without restriction.
Sentiment Analysis	Facilities which enable analysis of a body of data/texts to quantify certain emotional conditions (among other things). Can be in the form of a list of words or terms with a corresponding score to be able to assess and score text in various domains: health, social media, survey responses, etc. For example, if a patient who has had surgery and has left the hospital keeps a diary, the narrative can be scored automatically. A negative score is given to certain terms such as ‘very painful’, and ‘immobile’; and positive scores are given to terms such as ‘less painful’, ‘more flexible’ and ‘started walking’.
Speech-to-Text Engine	Software that turns the spoken word into written text.
Stemmer	Software or script which cuts the end of words to reveal the stem. For example, the Welsh verb datblygu (to develop): ‘datblygodd’, ‘datblygais’, ‘datblygiad’, ‘datblygiadau’ would be cut to the stem, which is ‘datblyg-’. Additional work is needed to deal with irregular verbs. While the lemmatizer deals with syntax, the stemmer reveals the meaning of words in sentences, the semantics. A useful tool in the development of machine learning and artificial intelligence in Welsh.
Stopwords	Functional words, such as ‘and’, ‘the’, ‘or’, which do not add to the themes or meaning of a text. A list of stopwords is used to filter the text and leave behind only the keywords.
Term Extraction	A script which studies the frequency of word order and highlights pairs or sets of words that are likely to be terms. It can identify and tag terms within passages automatically, meaning that the terms, not the individual words, are analysed/translated.
Translation Memory Software	Simply put, software that *remembers* previous translations and offers them to the human translator. Translation memory programs create a table of translations, by segment, from one language to another. Such systems help to ensure consistency and avoid the need to translate the same segment more than once. (Such systems are different from machine translation systems, although one can be used within the other). Translation memory is an example of a computer-assisted translation (CAT) tool.

Term	Definition
Translation Technology	Software that enables automatic or partially automatic translation. Such technology can be in the form of translation memory software which inserts segments from translation memories (see above), where there is a likelihood that a segment has been previously translated. It can also be in the form of automatic translation when there is no equivalent sentence already in a specific translation memory, or when an approximate(gist) translation, is required.
User Experience (UX)	The experience the user will have when using a particular technological system.
Voice Banking	The process of storing recordings of people reading a script or talking naturally. Software can then be used to cut the sound created into a series of pieces. These pieces could be used to create, for example, text-to-speech software (i.e. personal synthetic voice).
Wikidata	An open database of knowledge that can be read and edited by both humans and machines. It acts as the central storage for the structured data of Wikipedia.
Word Sense Disambiguation	A collection of natural language resources which use probability techniques and the context of words in relation to other words to improve the ‘understanding’ by machines of Welsh content. Disambiguation also applies to lists of terms and other systems that help with using the correct terms in the correct context.

Appendix 1: Welsh language infrastructural building blocks

This *Action Plan* emphasises the need to maintain up-to-date linguistic infrastructure to enable us to realise the vision of *Cymraeg 2050*. Infrastructure consists of many different elements, e.g. corpora, dictionaries and terminological resources. The following technical appendix is not exhaustive. In it, we outline some of the necessary technical building blocks, to develop an appropriate presence for Welsh in language technology (the item numbering follows on from the main list of components above).

N°	What	Why	How	Deliverable	Risks of not undertaking this work
19	Aligned Welsh/English parallel text published under an appropriate licence.	Such data is required to train machine translation systems	By creating/acquiring/issuing parallel texts in various linguistic registers. To be done through discussion with institutions with a view to ensuring that they contribute text or speech to the corpora.	Texts available under an appropriate licence.	Lack of training data for machine translation systems.
20	Stemmer	A stemmer cuts the end off words to reveal the root. For example, the Welsh language verb datblygu (develop): 'datblygodd', 'datblygais', 'datblygiad', 'datblygiadau' would be cut to the stem, which is 'datblyg-'. Additional work is needed to deal with irregular verbs. While a lemmatizer deals with syntax, the stemmer reveals the meaning of words in sentences, the semantics. It is a useful tool in the development of machine learning and artificial understanding in Welsh.	Develop a set of rules and a list of exceptions.	Software available under an appropriate licence.	Delays in terms of the development of machine learning and artificial understanding in Welsh.
21	Parsers: dependency parser, constituency parser.	To reveal semantic meanings of sentences for machine translation, conversational AI , etc.	Computer analysis of sentences to split into individual parts and create a tree that reveals the grammar of the sentence and the conceptual relationships between words.	Software available under an appropriate licence.	Errors when recording, transcribing and translating words.
22	Word and Term Embeddings	Welsh word sense disambiguation. Probability techniques and the context of words in relation to other words are used to improve the 'understanding' of Welsh content by machines.	Learning algorithms that use huge corpora and word embeddings vectors within the language model to consider the word in its context and offer inferred meaning on the basis of probability.	Embeddings available for Integration into translation engines, speech-to-	Errors when recording, transcribing and translating words.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
				text technology, translation automation systems etc.	
23	Term Extraction	To automatically identify terms within texts and store them in a glossary. Defining and treating terms consistently is one element of reducing the risk of translation errors.	A computer script which studies word order frequency and highlights pairs or sets of words which are likely to be terms.	Software available under appropriate license.	Errors when translating words automatically and failure to notice inconsistencies in the use of terminology.
24	Welsh Language Named Entities	Certain entities are useful in preserving words and terminology and ensuring that they are treated as one entity. In machine translation, named entities are used to 'protect' units of meaning from being treated separately e.g. so that a person's name such as 'Dr Smith' is not translated literally as the Welsh 'blacksmith' 'Dr Gof'. When analysing text, a set of job titles such as 'Chief Executive', 'Director ', 'investigating officer, etc. can be identified as named entities so that the terms are treated as one concept, in this case part of the family of job titles.	Analysis of lists and creation of new lists of different categories, such as names of people and things.	Relevant lists available under an appropriate licence.	Incorrect translation, inability to use Welsh to mine texts for concepts, probability and data analysis.
25	Welsh language sentiment analysis scores for broad domains, names, idioms etc.	For use in the health sector and marketing, among others.	Analysis of corpora of text from relevant domains to extract Welsh terms. With specialist assistance, a score can be allotted to each of the relevant terms.	Sets of scores available under appropriate licence	Only text in other languages can be analysed and assessed automatically.

N°	What	Why	How	Deliverable	Risks of not undertaking this work
26	List of Welsh Stopwords under an appropriate licence.	Stopwords are words that are filtered out before or after natural language processing in computerised form, to reduce processing time and prioritise important words.	Use of word frequency lexicon to recognise linking words which are not essential to the meaning of sentences.	Relevant lists available under appropriate licence.	Automatic processes will slow down because they are dealing with less important words which have to be filtered at the end.
27	Welsh language WordNet	To reveal the true meaning of ambiguous words, to improve machine translation, speech-to-text, etc. A WordNet examines relationships between words and synonyms and helps to refine the meanings, e.g. when translating in applications which use conversational AI.	This has been initiated by Cardiff University with financial assistance from the Welsh Government.	Welsh WordNet available under an appropriate licence.	Delays in terms of machine learning and artificial understanding in Welsh.

Appendix 2: Background

In *A Living Language: A Language for Living* (the Welsh language Strategy for 2012-17), technology and digital media was a key focus. To achieve the aims of that strategy, we worked with the Welsh Language Partnership Council and other specialist groups and stakeholders. This work led to an ICT *Action Plan* for the Welsh language in 2013. This stated the Welsh Government's commitment to stimulate developments in this area.

The purpose of that *Action Plan* was to build on work done in this area by the former Welsh Language Board. One of the Board's most notable successes was the work undertaken with Microsoft Corporation to secure Welsh-language interfaces for Windows, Office and SharePoint. The plan also emphasised the importance of the localisation, content creation and development work that had been done, and that is still being done by a small, enthusiastic community of volunteers.

These are the five key themes from the previous 2013 *Action Plan*:

- Marketing and raising awareness
- Motivating the key technology companies to increase Welsh-language provision
- Encouraging the development of new Welsh-language software applications and digital services
- Stimulating the creation, sharing and consumption of Welsh-language digital content
- Supporting good practice across the public, private and third sectors

In implementing that *Action Plan*, the Welsh Government's main activities over that period were:

- Funding Welsh-language digital media and developments through our Welsh-language Technology and Digital Media fund.
- Encouraging the use of Welsh-language computer interfaces via the use of 'How to...' videos.
- Working with large technology companies.
- Developing networks in Wales and beyond (e.g. The Minister's Technology Board, META-NET).
- Providing advice to Welsh-language organisations in the third sector on their use of digital media.
- App of the Week
- Working with other departments of Welsh Government in the context of the Welsh language and technology.

To support the *Action Plan*, a £750,000 fund was created over three years. Its objectives were to support activities as follows:

- Raising awareness of Welsh-language software, programmes, online services, resources to create content and digital content or promote the use of these resources.
- Supporting the work of developing Welsh-language software and digital online services.
- Increasing the volume of Welsh-language digital content available online.

The fund supported a diverse range of projects, which roughly come under these headings:

- Developing language technology infrastructure (e.g. spoken command recognition) (5 projects).
- Digital content (e.g. apps, websites) (8 projects)
- Digital content enablers (e.g. devices and blog themes) (4 projects)
- Develop digital skills (e.g. training sessions on creating content, coding clubs) (4 projects);
- Digital services that link with the promotion of the Welsh language (e.g. apps, events) (5 projects).

The Welsh Government's *Cymraeg 2050: A Million Welsh Speakers*[13] *Strategy* notes that we will "Ensure that the Welsh language is at the heart of innovation in digital technology to enable the use of Welsh in all digital contexts." Two other strategic documents have led to the creation of the current *Welsh Language Technology Action Plan*.

***Cymraeg 2050*: Work Programme 2017-2021[14] relevant sections**	***Cymraeg 2050* 2018-19 Action Plan[15]**
Section 12. Digital technology	
Our aim: ensure that the Welsh language is at the heart of innovation in digital technology to enable the use of Welsh in all digital contexts.	
By 2021 we will do the following.	
12.1 Invest more in research and innovation in language technologies to facilitate use of Welsh in the digital age	We will form a detailed Technology and Welsh Language Action Plan through the Minister for Welsh Language and Lifelong Learning's Technology Board, identifying all necessary steps to normalise and support the Welsh language in technology.
12.2 Explore investment opportunities, collaboration, the sharing of resources and techniques to support our technological infrastructure (computer-aided translation, artificial intelligence (AI) technology, voice recognition, etc.) so that the Welsh language can be used on electronic devices.	12.2 The process of drafting the Technology and Welsh Language Action Plan and of implementing it means that we will examine opportunities to invest, collaborate and share resources and techniques.

[13] https://gov.wales/docs/dcells/publications/170711-welsh-language-strategy-eng.pdf
[14] https://gov.wales/docs/dcells/publications/170711-cymraeg-2050-work-programme-eng-v2.pdf
[15] https://gov.wales/topics/welshlanguage/welsh-language-strategy-and-policies/welsh-language-policies-upto-2017/welsh-language-strategy-action-plan/?lang=en

***Cymraeg 2050*: Work Programme 2017-2021[14] relevant sections**	***Cymraeg 2050* 2018-19 Action Plan[15]**
12.3 Support the development of new digital bilingual resources for use in schools, the workplace, and socially.	12.3 We will identify needs and develop a commissioning programme for schools to meet the needs of the curriculum and qualifications. The priority in the workplace will be machine translation and promoting systems that help translators to automate their processes and share their translation memories. Socially, the work of developing a speech-to-text system will begin, which will mean that Welsh-speakers do not need to type commands on their phones, tablets and computers.
12.4 Work on motivating the main technology companies to increase the Welsh-language provision offered.	12.4 Dialogue will be facilitated by developing the kind of Welsh-language technology infrastructure that is needed by the major companies to be able to offer support to Welsh-speakers. The challenge will be to convince them of the business case for doing so.
12.5 Ensure that our grant recipients and organisations promoting the Welsh language are using technology well, including data systems, internal communication, social media and marketing tools.	12.5 In 2018–-19, in allocating the grant for the promotion of the use of the Welsh language, we will ensure that technology and marketing areas are included. We will also share information with establishments that promote the language about new developments in these areas.
12.6 Support efforts to increase the number of Welsh-language Wikipedia pages.	12.6 It is intended to continue to encourage communities of Welsh-language Wikipedia volunteers by offering financial support for workshops and the creation of automated Welsh articles. The Minister for Lifelong Learning and Welsh Language will address the Celtic Knot conference at the National Library in Aberystwyth on 5 July 2018.

Appendix 3: Contributors

The Welsh Government's Welsh Technology Board:

- Minister for Lifelong Learning and Welsh Language, Eluned Morgan AM (Chair)
- Andrew Hawke, University of Wales Dictionary
- Delyth Prys, Bangor University's Language Technologies Unit
- Dewi Bryn Jones, Bangor University's Language Technologies Unit
- Dr Dafydd Trystan, Coleg Cymraeg Cenedlaethol
- Dr Daniel Cunliffe, University of South Wales
- Dr Eleri James, Welsh Language Commissioner
- Helgard Krause, Welsh Books Council
- Dr Rhodri ap Dyfrig, S4C
- Gareth Morlais, Welsh Government
- Glyn Rogers, Ysgol Gyfun Gwynllyw
- Huw Ynyr, Gwynedd Council
- Illtud Daniel, National Library of Wales
- Dr Jeremy Evas, Welsh Government
- Owen Derbyshire, Properr, Welsh Language Partnership Council
- Rhys Evans, BBC Cymru Wales
- Sara Huws, National Museum of Wales
- Sarah Dafydd, National Assembly For Wales
- Steve Morris, Swansea University

We also gratefully acknowledge the additional input provided by:

- Prof Steve Renals, Centre for Speech Technology Research, School of Informatics, Edinburgh University.
- Professor Alun Preece, Deputy Head of School of Computer Science and Informatics, Cardiff University.
- Hacio'r Iaith.