Llywodraeth Cymru
Welsh Government

Welsh language technology action plan

Progress report 2020

Welsh language technology action plan: Progress report 2020

Audience

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

Overview

This report reviews progress with work packages of the Welsh Government's *Welsh language technology action plan* between its October 2018 publication and the end of 2020. The *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 action plan* (2018); *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.

OGL © Crown copyright 2020 WG41736 Digital ISBN 978 1 80082 629 8

Contents

Ministerial foreword 3

Introduction 4

What we’ve done so far 5

What we’re doing now 7

Background 10

Making sure all this work gets used as much as possible 11

How much has this all cost? 12

Work with the major technology companies 13

Progress of individual Work Packages 14

Glossary........ 25

Ministerial foreword

When I published our Welsh Language Technology Action Plan in October 2018, I said we need to grasp opportunities and tackle technological challenges by trying to anticipate wider technological developments and set a direction for technology and work in the Welsh language.

I set out 27 work packages, with the emphasis on speech, translation and artificial intelligence. As this work is publicly funded, I was determined that the products created under each work package would be available free of charge, for everyone to use and adapt. And that's why there are so many of them available for you to download today under an open license.

2020 has been a difficult year. During emergencies like COVID-19, technology can help us to deliver important messages quickly. This is why I've been flexible in re-prioritizing aspects of the Action Plan, to respond to fast-changing needs during the pandemic.

I brought forward some release dates, one of which is Cysgliad, the Welsh spelling and grammar checker. This is now available from Bangor University free of charge for individuals, all schools and small businesses to use. I felt this was important for school learners and their parents during the first lockdown, with so many children learning independently at home while schools were closed. Sometimes, there'd be no adult in the house who spoke Welsh to help them with their Welsh writing. So Cysgliad can make a real difference to them and to many others.

I've also asked Bangor University to prioritize automatic subtitling of Welsh videos. The requests from universities to caption Welsh lectures on video are increasing. This work was timetabled for later, but as so much teaching is now online, I've asked for this work to be brought forward.

Technology has allowed events that would have been cancelled in 2020 to be held virtually. Elements of the Urdd and National Eisteddfodau moved online, and online video meetings have become commonplace during the pandemic. This has presented a number of challenges for the Welsh language, as simultaneous translation is still not available in all packages. If it were, we could use more Welsh. You can read more on this later.

Cardiff University's School of Computer Science and Informatics is currently developing work on word embeddings. This will improve the way computers can understand the meaning of Welsh text and the intent of the users. This has led to the possibility of creating new games to help those learning Welsh.

It is only by working together that we've able to begin to realize the Plan's objectives and contribute to doubling the daily use of Welsh by 2050. I know that technology develops rapidly. I'm keen for the Welsh language to move with those developments. That will be the case as we implement the rest of this Plan, moving next to exciting developments in computer-assisted translation.

- Eluned Morgan MS - Minister for Mental Health, Wellbeing and Welsh Language.

Introduction

The Welsh Government's *Welsh Language Technology Action Plan* (henceforth 'the *Plan*') was announced via oral statement in the Senedd in October 2018.

In launching the *Plan*, the Minister for International Relations and the Welsh Language (now Minister for Mental Health, Wellbeing and Welsh Language) said she wanted to see people using Welsh language technology and to make sure that Welsh is at the heart of innovation in digital technology. The aim would be to make it possible to use Welsh in all digital contexts. This Report sets out progress to date towards that aim.

The Minister has stated she believes technology is a 'game changer' in language planning and is something that drives the Welsh language policy agenda forward. But to change the game, we need the right components. The aim and philosophy of the *Plan* is that those components are created and made available under a suitable open licence, so everyone can use them time and again. As you read on, we hope you'll see that philosophy being implemented in all the elements we're funding and have funded in carrying out the *Plan* and its work packages.

The three specific infrastructural areas the *Plan* addresses are: Welsh language speech technology, computer-assisted translation, and conversational Artificial Intelligence.

What we’ve done so far

We’ve made significant progress in implementing the *Plan*’s work packages. At the time of drafting this document, we’re implementing or have completed 19 of the work packages from the total of 27. We’ve put plans in place to implement the others. See below for detailed information on the progress made for each work package, including links so you can - free of charge - download components already created.

Here are a few highlights of the work already completed under the *Plan* with our funding and/or support:

- We’ve made Cysgliad, the Welsh language grammar and spelling checker and a series of dictionaries available free of charge to individuals, organisations with ten or fewer employees, and to all schools in Wales (work package 17).
    - We're aware that confidence in their Welsh is a problem for many as they create content in Welsh.
    - releasing Cysgliad for free will contribute to building that confidence with the aim of increasing the amount of Welsh that is written.
    - several thousand copies of the free package have already been downloaded. 5,032 downloads as of 15 December 2020.
- Bangor University has improved the virtual assistant Macsen, This has the potential to offer future benefits for Welsh-speaking people who are using digital assistants more and more for accessibility purposes.
    - to ‘understand’ spoken Welsh, the university has created acoustic and linguistic models for Macsen to identify the 2,500 Welsh words and those 500 English language words most commonly used in spoken Welsh
    - Macsen uses these English words in order to deal with code switching between Welsh and English.
    - This capability has been used in new apps for both iOS and Android mobile devices and in a Windows 10 Office add-in for Welsh transcription.
- Bangor University has also released an improved version of the Welsh wordlist called Hunspell, released a Neural Parts of Speech Tagger and new Language Normalisation Tools.
- Google for Education and Adobe Spark are available in Welsh (Work package 7).
- From November 2020, the default interface language for Microsoft Office 365 changed from English to Welsh for Learners in schools that teach through the medium of Welsh. This affected 78,086 learners in 379 schools. (Work package 7).
- Cardiff University has been researching Welsh word embeddings and the vectors they produce to improve the way computers can understand the meaning of Welsh text and, in so doing, users’ intentions (Work package 22).
- We are Service Works Ltd (formerly Satori Lab Ltd) is further developing the Welsh version of Open Streetmap (Work package 18).
- As a result of our school coding campaign *Cracking the Code*, a large number of coding learning resources are available bilingually. This has the potential to open up coding, digital transformation and other digital professions to a wider audience (Work package 8).

- Among the activities undertaken with our grant by the National Library of Wales to increase the use of the Welsh language through the use of Wikipedia were:
    - the creation of 50,000 bilingual Wikidata items for books, authors and all publishers in Wales with 100 or more publications.
    - the publication of 500 Wikipedia articles about female authors, using open data.
    - public workshops where volunteers wrote and edited Welsh Wikipedia articles about literary subjects.
    - with the collaboration of Menter Môn, developing and holding eight events in schools in north Wales, working with teachers and learners of different ages to create Wikipedia articles about Welsh literature. The emphasis was on titles being studied by the learners themselves
    - all of the social and cultural capital created through this community work (Work Package 15).

What we're doing now

The main new work in progress under the *Plan* is the work we're funding Bangor University to undertake. During the financial year 2020-21, we've given a grant of £347,950 to the University. The grant was been made conditional on Cysgliad being released free of charge (as noted in several parts of this Report). The work will also see the creation of new Welsh language text-to-speech voices and new skills for the virtual assistant Macsen. The following table details this and other work.

Table: Components that will have been created under the *Plan* by the end of 2020-21

Work Package	Work undertaken	Developer
1	New Welsh speech-to-text transcription software.	Bangor University
1	New Welsh-language speech-to-text acoustic Model.	Bangor University
2	Corpus of anonymised Welsh text.	Bangor University
2	A new skill for Welsh virtual assistant Macsen.	Bangor University
2	Update to the Welsh language model of the virtual assistant Macsen.	Bangor University
5	A revised version of the Welsh personal voice banking application *Lleisiwr*, which produces a personal and unique text-to-speech voice for any user. This is valuable to Welsh speakers who have a health condition, which could threaten their own voice.	Bangor University
5	Four new bilingual (Welsh-English) text-to-speech voices.	Bangor University
6	New automated Welsh language quiz software for Welsh learners. The quiz can be inserted into various websites and questions are offered on the basis of word embeddings. This is an example of a development for Welsh learners arising from Cardiff University's language technology infrastructure. Word embeddings will enable hundreds of questions to be created that will help learners.	Cardiff University
7	Default language of pupils interface language within Hwb's Office 365 tenant for Welsh medium schools to be Welsh. This will enable many thousands of learners to have an enhanced Welsh language user experience. We are engaging with tech companies such to further promote Welsh language localisation.	Hwb website (Welsh Government)

Work Package	Work undertaken	Developer
8	Welsh language coding resources Cracking The Code on our education website Hwb.	Code Club, Technocamps and others.
9	Welsh language Touch Typing programme to help blind and visually impaired children to learn to type.	Welsh Government
10	A lexicon of Welsh language words.	Bangor University
14	Compilation and publication of a list of Welsh language software.	Welsh Government
15	The focus for our support to the Wikipedia Welsh language community activity this year is photography with a project called #Wici-Pics. Nine online #Wici-Pics workshops or events will be held and Welsh data for 6,426 chapels in Wales will be added to Wikidata in 2020-21.	The National Library of Wales assisted by Menter Iaith Môn
15	#WikiAddysg: new Welsh Wikipedia articles about the subject History to ensure suitable resources are available for the new curriculum at school.	The National Library of Wales assisted by Menter Iaith Môn
16	Corpus of Welsh sentences tagged with parts of speech.	Bangor University
17	Update to the Welsh Hunspell word list.	Bangor University
18	Improvements to OpenStreetMap Wales to show more place names, streets and geographical features such as lakes in Welsh on an interactive map.	We Are Service Works Ltd.
19	107 translation memories had been released free of charge for use under open licence on the Byd Term Cymru website at the time of compiling this document. Translators can load and use these memories into their own translation systems for reuse.	Welsh Government
21	A Welsh part of speech tagger developed using neural networks.	Bangor University
21	Welsh language text manipulation scripts for mutations, plural forms, etc.	Bangor University
21	Normaliser to pre-process Welsh language text.	Bangor University

Work Package	Work undertaken	Developer
22	New vectors for Welsh language text models.	Bangor University
22	An academic paper on the process of using cross-linguistic embeddings to create new natural language processing tools for the Welsh language based on English language data. Implications and lessons for other smaller languages throughout the world are discussed.	Cardiff University
25	A new automatic sentiment/opinion analysis tool for Welsh language texts, developed using cross-linguistic embeddings.	Cardiff University

Background

Here's some background about how we went about creating and implementing the *Plan* and the choosing the components produced: quality control and independent peer review.

There are a number of ways of evaluating the *Plan’s* outputs and progress. This section sets these out. Quality is paramount for products and components we create, as is our wish for them to be released under a suitable open licence and to be widely adapted and adopted. With this in mind, we’ve put these measures in place:

(1) The *Plan* was created in conjunction with the Welsh Language Technology Board, chaired by the (then) Minister for International Relations and the Welsh Language. This co-creative approach was a way of ensuring the input of a wide range of stakeholders who had the skills and experience of working in the world of language technology and/or of its implementation. A list of the members of that board is available at the end of the *Plan*.

(2) In addition to this, having published and begun implementing the *Plan*, we’ve established a formal ‘peer review’ mechanism for the evaluation of work that we may fund. Several internal and external experts undertook these reviews. The aim of this was to challenge us and our potential contractors constructively and to help us achieve the best possible standard of work.

(3) As well as this, there will be further peer reviews of new components created under our new grants (as noted, the main new grant is currently for work being undertaken by Bangor University).

   i. These reviews will be carried out by an independent computational linguist, unconnected to the project.

   ii. We’re doing this to ensure that all components are of the highest quality, and that they conform to relevant international standards, (so that they can be used by as many systems and organisations as possible).

   iii. We stress that the philosophy of the *Plan* is that components we fund should be available free of charge to everyone under a suitable open licence, as far as possible.

(4) We also spent a substantial amount of time reviewing potential work with a software intellectual property lawyer to ensure our ‘free’ software philosophy was implemented.

Making sure all this work gets used as much as possible

Making sure companies, organisations and people make use of what's created is key, as is awareness and ownership of what's in the pipeline.

The main purpose of funding software components is so that they can be used to facilitate and increase the use of Welsh. For that to happen, as well as *existing*, the components must be *used* in software. For them to be used, people/companies/organizations must know about their existence, appreciate the need for Welsh components, and include them in their products.

Please note that this paper doesn't necessarily discuss components that the public themselves use directly. What the *Plan* creates are infrastructural components that, when implemented in other programmes/apps/products, will enable the Welsh language to be used in situations where Welsh speakers can currently only use English. Smart personal assistants, (or smart loudspeakers such as Alexa, Siri, Google Assistant)), are one possible example of the type of software made by several manufacturers that could use the components we create.

Here's how we're aiming to ensure the use of the components that we create and how we are building ownership of the philosophy and outputs of the *Plan*:

(1) We made it a requirement on Bangor University to convene a network of independent experts (by open invitation) to ensure external expert inspection of the work the University undertakes with our financial support. Welsh Government officials will also be formally involved in the overview of each work stream. This is in addition to the peer review noted above.

(2) Also, as part of Bangor University's grant, we've made it a requirement that the University must create and implement a plan to ensure the components they develop are actually used by external organisations. The Welsh Government similarly emphasises these products in the advice it provides to external organisations and internal colleagues, and as we utilise the procurement system for the purposes of driving the Welsh language software market.

(3) The *Welsh Language Technology Action Plan* is the subject of scrutiny by internal Welsh Government committees, including the Welsh Government's Digital and Data Officials Group.

(4) In addition, officials carry out work on the *Plan* in various Welsh Government departments. This is part of the work to mainstream the Welsh language across all policy areas in line with the first Minister's leadership manifesto. Procurement is one example of this, as is *Cymraeg. It belongs to us all*, the Welsh Government's policy for the Welsh language internally.

How much has this all cost?

Between its launch in 2018 and the end of the 2020-21 financial year we will have spent £651,000 on the implementation of the *Plan*. This document details the work undertaken with that expenditure. The *Plan* leads us beyond the current Senedd, and will be fed into the next *Cymraeg 2050* five-year plan*.*

For 2020-21 we've awarded four technology grants:

- £348k: Language Technologies Unit, Canolfan Bedwyr (Bangor University) to work on the *Plan*'s work packages. We detail this grant in a number of places in this paper.
- £90k: Cardiff University: To develop automatic sentiment analysis of Welsh language texts using innovative cross-lingual word embeddings.
- £15k: The National Library of Wales: to crowdsource Welsh photos and stories in Welsh in a project called #Wici-Pics. This is in conjunction with Menter Iaith Môn.
- £10k: Mapping Wales (The Satori Lab (now We Are Service Works) Ltd.) To continue to develop interactive maps displaying Welsh-language place names and interface.

It should also be noted that further expenditure on technology and the Welsh language is made by the National Centre for Learning Welsh and funding streams beyond the portfolio of the Minister for Mental Health, Wellbeing and Welsh Language (e.g. resources in Welsh for those with additional learning needs, Cracking the Code's bilingual resources, etc.).

Work with the major technology companies

Microsoft is a good example of how the public sector has worked with international technology companies. For nearly 20 years, Microsoft’s Welsh language provision has increased, from a spellchecker, to a series of user interfaces (Windows, Office, SharePoint) and other tools that enable the use of Welsh (the Welsh Language Board began this work and it was transferred to the Welsh Government on the abolition of the Board in 2012). These are all available free of charge and have been created without financial expense to the state. We have a constructive relationship and an ongoing dialogue with Microsoft about many matters, Welsh language provision being just one.

We recognised at the beginning of coronavirus lockdown that Microsoft software such as Microsoft Teams, facilitated the continuation of organisations’ work. We also realised it was a problem to hold these video meetings bilingually via simultaneous human interpretation (other software offers this facility). The Minister wrote to Microsoft to make the case for introducing the ability to offer simultaneous human interpretation to its Teams software (which will eventually replace Skype for Business) and shared her letter on her Twitter feed. If this facility is developed, there will be an increase in the use of small languages throughout the world (it also applies of course to all major international organizations that are multilingual).

As set out elsewhere in this paper, we’ve worked with Google to ensure that Google for Education, including Google Classroom is available in Welsh. We’re in discussions with a number of other organisations such as Adobe, which has released a Welsh version of Adobe Spark. From November 2020, we switched the default interface language of Microsoft Office 365 from English to Welsh for 78,086 learners in 379 schools which teach through the medium of Welsh.

We think our policy of funding development which is shared under a permissive, open licence helps to make it as easy as possible for large companies to add the Welsh language to their own services and makes it easier to make a business case for a relatively small market.

We’ve recently been holding meetings with a number of large international companies and have provided advice to those planning to add Welsh to their products and services. E.g. we can offer the components and data we create, including translation memory data, so Welsh can be added to companies’ products and services.

Our work in procurement, together with our openly-licenced free software philosophy will also facilitate the future development of bilingual software by all manufacturers.

Progress of individual Work Packages

Description and *Action Plan* Work Package number	**What's been done/What's being done**
WP.1. Welsh-language speech-to-text facilities and components.	We're funding Bangor University to develop Welsh speech-to-text components. As a result of this work, it's now possible to identify and transcribe the 2,500 words most commonly uttered when speaking Welsh. The facility can also deal with the English words which are most commonly used by people speaking Welsh when code switching. A talk and type extension or add-on has been created for Microsoft Office on Windows 10. When examining how such technology can be of use to the Welsh language, we hold discussions with a number of organisations, e.g. to see how Welsh audio recordings can be transcribed automatically. Bangor University is looking at ways of automating subtitles on Welsh videos.
WP.2. Welsh-language machine learning and conversational Artificial Intelligence.	We're funding Bangor University to further develop a virtual assistant (Macsen). It's now possible to speak with Macsen in a natural Welsh to ask it to complete tasks or to request information. A Macsen app has been created for Apple iOS and Android. We've also been discussing with private companies how these skills could be incorporated into their smart devices.
WP.3. Welsh language audio corpora for annotation, to train speech-to-text and other technologies.	We've already funded 'Paldaruo', a crowdsourcing app to collect Welsh voice data. It's used to train technologies and for academic research. We've also supported the campaign to get as many people as possible to contribute to the Mozilla Common Voice project https://voice.mozilla.org/cy (this link only works on Firefox and Chrome browsers) by sponsoring marketing work on social media and by using our own social media channels. Common Voice is a crowdsourcing programme that collects voice data for many of the world's languages, with the aim of facilitating speech technology in those languages. The corpora created through Common Voice are available for downloading and use under open licence. 94 hours of Welsh language voice recordings are available for download at the time of writing. Through the European KESS II programme we've funded PhD research into Welsh language speech technology, by analysing the speech and subtitles of S4C programmes. The impact of differences in local accents on the accuracy of transcribing Welsh voices has been an important aspect of this research.

Description and *Action Plan* Work Package number	**What's been done/What's being done**
WP.4. Improve user experience of Welsh language technology through behavioural economic techniques.	When drafting the Welsh Language Technology Action Plan we heard that some people struggle to access Welsh-language technological services despite them being available. Consequently the use of some of them is low. This is often due to the way the language choice on that service is architectured (it expects the Welsh speaker to *look* for the service rather than a given organisation offering it proactively). Work in this area is also linked to work packages 7 and 13 and also our policy for working bilingually internally in the Welsh Government *Cymraeg. It belongs to us all*, in which behavioural economic techniques are used to increase the use of Welsh in technology. This work is detailed under each specific work package section. There's another example of this in our work to reduce the friction in finding the Welsh interface for thousands of Welsh-speaking learners in school.
WP.5. Frameworks to personalise text-to-speech and bank individual voices.	The foundations of this work were laid as we funded a pilot version of the *Lleisiwr* voice banking application by Bangor University. This allows individuals to record scripts that are then turned into TTS (text-to-speech) versions of that person's Welsh and English speaking voice. The university has since developed its voice banking even further as part of our work in the wider text-to-speech world, and Lleisiwr 2 is now available. From our discussions with stakeholders, we understood that new Welsh-language voices were needed. Creating these is part of our work, although it has been delayed slightly because of COVID-19. By the end of 2020-21, we hope four new Welsh voices will be available for use under open licence (one male and one female voice with a southern Welsh accent, and the same number as a northern Welsh accent).
WP.6. Interactive content and software for Welsh learners.	The Centre for Welsh Language Learning has already planned to deliver much of its training online and it's fair to say that the Welsh learning sector had responded well to the COVID-19 situation by its use of the available technology to facilitate its work. For example, all face-to-face community lessons that began in September 2019 or January 2020 are now being maintained on-line. This has enabled all learners to continue to learn Welsh as well as to bring other benefits e.g. using technology to keep in touch with other people during the lockdown period. The Centre has also launched a new blended learning course which is being run online so learners can work independently

Description and *Action Plan* Work Package number	**What's been done/What's being done**
	and have sessions with tutors. Over 7,000 new learners registered an interest in the course, and 1,300 learners have now started the course in 89 different classes. The centre has started to run "Welsh at home" daily lessons on Facebook. Over 1,000 people look at the lesson every day. This helps to bring Welsh into homes while schools are closed. Cardiff University is developing new games for Welsh learners. This comes as a result of its work on natural language processing and the university is in collaborating with the Centre for Welsh Language Learning on this. In May 2020 the Centre announced a new partnership with *Say Something in Welsh* (SSIW). The centre and SSIW will share resources and It'll be possible for learners of either provider to enjoy a discount for access to the lessons of the other. This is a step towards creating a single community for Welsh learners and increasing access to courses of different kinds. Learners will also have the opportunity to join the virtual community of SSIW, where it's possible to converse with other learners. There will be further opportunities for both organisations to share good practice and to share the benefits of technological developments. Future plans for the Centre include live question and answer sessions on Facebook, sessions for parents and carers of young children and the introduction of blended learning lessons at other levels. The Centre will also be relaunching its programme '*Siarad*', which matches Welsh speakers with learners, with the aim of helping to develop confidence in using the Welsh language.
WP.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.	This work has been done for Welsh-medium schools. It involves the Hwb Office 365 interface. Hwb is one of the largest single education tenancies in Microsoft around the world, with over 550k users. Microsoft Office 365 was already available in Welsh for school learners in Wales through Hwb, but each individual user had to make a decision of his or her own accord to choose to switch their language interface from English. This Work Package has ensured that learners in Welsh-medium schools are presented with the Welsh language interface on Office 365 as a default, without them having to do anything.

Description and *Action Plan* Work Package number	**What's been done/What's being done**
	This is an example of our work using behavioural economic techniques noted in Work Package 4 above (defaulting being one the cornerstones of the MINDSPACE behaviour change model, and its EAST sub-model). One of the aims of defaulting the Hwb interface and services to Welsh for learners in Welsh medium schools is to make it 'EAST' (Easy, Attractive, Social, Timely) for them to use the language, i.e. remove the 'friction' (the difficulty) from using the Welsh language. Since the publication of the *Plan*, we've also worked together with Google to ensure Google for Education is available on Hwb in Welsh, and so too is Adobe Spark.
WP.8. Promote Welsh language technology and coding resources to teachers and children and others.	We've funded Welsh language coding resources through several sources (for example, Cracking the Code resources are available bilingually on our Hwb website. Technocamps (Technology workshop providers) reported that the number of teachers who had attended their course to train teachers to teach coding through the medium of Welsh had grown from 40 in the year 2017-18 to 49 in 2018-19. The latest data about use of the resources will soon be available. We're also in discussions with other external suppliers with a view to providing practical support for them to increase their training provision.
WP.9. Create and/or develop facilities to assist Welsh speakers with additional learning and/or accessibility requirements.	There's a need to identify the gaps in additional learning needs (ALN) provision in Welsh. E.g. while a learner with dyslexia who writes an essay in English can use a speech-to-text resource to dictate English language content, the same resource is not available in Welsh. The speech-to-text infrastructure work that we're funding and share under open licence in Work Package 1 will therefore contribute to enabling education companies to create such a resource. We've held discussions with ALN experts, e.g. Teachers who teach blind and visually impaired learners. On this basis, we've commissioned a new Welsh touch typing Learning Resource. The resources released as a result of our work will help with aspects of digital accessibility in Welsh as well as assisting learners with ALN.

Description and *Action Plan* Work Package number	**What's been done/What's being done**
WP.10. The Workplace. Ensuring suitable English/Welsh and Welsh/English machine translation systems for different linguistic domains and registers.	We're funding and adopting in our own work, a number of streams of activity in this area. We're funding Cardiff University to develop a new type of machine translator that uses cross-lingual word embeddings to convert words and terms (this will help us to establish principles that can be used in other aspects of natural Welsh language processing, particularly those where there is much less Welsh language than English language data). In addition to machine translation, we freely and openly share translation memories. The parallel bilingual data in translation memories can also be fed into automatic machine translation products (see Work Package 11) to increase their productivity. We also hold discussions with external organisations, to facilitate their use of machine translation, and with some companies to assist them in developing machine translation products. IBM Watson Translate has introduced the ability to translate from Welsh to English. Cwm Taf Morgannwg University Health Board is the main partner. We welcome new Welsh language products like this from a large company like IBM.
WP.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.	Translation memories have a significant contribution to make to increasing the amount of Welsh available in the linguistic landscape. At the time of writing, we'd shared 107 translation memories on our BydTerm Cymru website, along with three large memories associated with new terms emerging as a result of COVID-19, all of which are available to be downloaded and used free of charge under an open licence. Translators in all organisations will be able to add these to their translation memory systems and take advantage of translation work already undertaken. We're also working to ensure that a translation automation plugin Works within our own Translation Memory system so that all the benefits it brings are available to our translators. We've also benefitted from the fact that we are now using 'SynchroTerm' terminology Software. This benefit has been especially pertinent during the COVID-19, period as it allows us to effectively mine large/urgent texts quickly and securely for our internal translators and contractors. We're also exploring what work can be done to facilitate the sharing of translation resources between different institutions,

Description and *Action Plan* Work Package number	**What's been done/What's being done**
	with a view to using what's shared as training data for automatic machine translation products which can be used by professional translators. Part of this work could also include 'scraping', tagging and aligning parallel Welsh and English language public texts, and then using this data to generate new translation memories and new domain-specific machine translation products.
WP.12. Modify, where relevant, procurement processes, so the Welsh language is a consideration in technology from the outset.	The principle that technology procured with Welsh public money should work smoothly in Welsh is important. To ensure this we've produced useful guidance to include in tender specifications and/or calls of different types that require technology. This guidance leads suppliers through an easy process to make sure the product they create do work bilingually. The guide also offer signposts for more detailed technical advice and, if that advice does not answer a question about technology and the Welsh language, we welcome enquiries to our officials. The piloting of this process has been paused because of COVID-19, but will shortly restart. The long-term ambition is that our procurement will drive the market for bilingual software in Wales. Of course, the principles of the work will apply to many multilingual situations around the world and we'll be sharing the work with our national and international partners.
WP.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.	We want people to be able to have digital services in their chosen language without each person having to ask for that service time and time again. To achieve this—i.e. to automate Welsh language services, that language preference data needs to be recorded. The language preference data should be recorded in a standard format which can be shared with other systems in the same institution and/or more widely (subject to relevant data regulations). The Welsh Government participated in a working group organised by the Welsh Language Commissioner on this theme (recording and sharing language choice) and we've responded positively to the recommendations of the Working Group's report. Following the work of the Working Group, the Welsh Government made a submission to WISB (Information Standards Board for Wales, which is part of the National Health Service Wales Informatics Service) [NWIS]) to carry out work on language choice data standards in NHS information systems. WISB accepted our proposal and the

Description and *Action Plan* Work Package number	**What's been done/What's being done**
	precise technical nature of the work is currently being scoped. We hope that this work with the health service will enable us to share case studies about language preference tracking systems that other organisations can use to create guidelines and systems for themselves. Facilitating language choice is also part of the work we're doing in procurement, detailed in Work Package 12. It should of course be noted automating language choice needs more than technical systems and standards. Those systems are only as good as the data that is fed into them. It'll therefore be necessary to train staff who operate these systems about the importance of entering language choice data, and how exactly service users should be asked about their preferred language. We should also be aware of the complex nature of individuals' language preferences that may vary from situation to situation, from institution to institution, and between services in the same institution.
WP.14. A list of Welsh language and bilingual ICT resources available in the workplace, also noting gaps in provision.	We're in the process of creating this list. It'll also be possible for the public to make enquiries about Welsh language software through our one stop shop service about the Welsh language (and its use in business), *Helo Blod*. We want the work we do in procurement to over time, to fill gaps in Welsh language software provision. All other aspects of the *Plan* also contribute to this: see, for example, the work that Bangor University will do to ensure the use of the components created with our financial support.
WP.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.	We're funding work that facilitates coordination of volunteer communities. These communities create original Welsh content on Wikipedia platforms. The number of Welsh-language articles on Wikipedia has increased from 35,807 in 2012 to 131,002 at the time of writing this document[1]. We funded the National Library of Wales in 2019-20 as part of the #WiciLlenyddiaeth project' "Llenyddiaeth" is Welsh for literature. As part of the project, workshops were held and content produced al about Welsh language books. Among the work undertaken was: • Creating 50,000 bilingual Wikidata items I about books, authors and all publishers in Wales with 100 or more publications.

[1] Source: https://stats.wikimedia.org/v2/#/cy.wikipedia.org/content/pages-to-date/normal|line|all|page_type~content|monthly

Description and *Action Plan* Work Package number	**What's been done/What's being done**
	• Publishing 500 Wikipedia articles about female authors, reusing open data. • Holding article writing and editing workshops which resulted in the publication of Welsh articles on Wikipedia about literary subjects. • With the co-operation of Menter Môn, developed and held eight events in schools in north Wales, working with teachers and learners of different ages to create Wikipedia articles about Welsh literature. The emphasis was on titles that were being studied by the learners themselves. This year we're funding the #Wici-Pics community photography project where the National Library of Wales and Menter Môn will work with schools and communities to crowdsource photography and Welsh stories from various parts of Wales. Nine #Wici-Pics online workshops or events will be held and Welsh data about 6,426 chapels in Wales will be added to Wikidata in 2020-21. The photos taken along with their Welsh language descriptions - will form a new digital collection at the National Library of Wales in Aberystwyth. #WikiAddysg has worked with subject experts to announce 71 new Welsh articles about history to ensure suitable resources are available for the new curriculum.
WP.16. Long-term support for the development of the linguistic infrastructure of the Welsh language, including corpora, lexicographical and terminological resources.	We're developing our policy for the linguistic infrastructure of the Welsh language. We'll be making announcements in due course.
WP.17. Welsh spellcheckers, grammar and mutation checkers available free of charge.	It's important that all possible resources are available to support the creation of Welsh language content and/or to boost the confidence of Welsh speakers to write in Welsh. So in May 2020, as part of our grant to Bangor University, Cysgliad (a Welsh spelling and grammar checker and a series of dictionaries) was released free of charge to individuals, all schools in Wales and organisations with fewer than ten employees. Our intention in doing this was that it would specifically help school children to create content and to do school work in Welsh at home during the lockdown period. Within two weeks of the announcement that the software was available for download free of charge, 5,032 copies had been downloaded by 15 December 2020.

Description and *Action Plan* Work Package number	**What's been done/What's being done**
WP.18. Interactive maps with Welsh language versions of place names that can be embedded within web pages.	We're funding a company called We're Service Works Ltd to work with OpenStreetMap volunteers in Wales to expand Mapio Cymru's Welsh language interactive map by: • Improvements to the functionality and loading speed of the map. Wales to show more Welsh place names and allow users to add and label more points of Welsh interest themselves: e.g. lakes, mountains, rivers, etc. • New methods for third parties to embed interactive Welsh maps on their own websites. For example, Mentrau Iaith Cymru have used this facility to show the location of a Welsh language pop concert on a map embedded on their own website. • Formation of new mapping communities. The implication of this is that the use of the Welsh language increases as groups of mapping enthusiasts work together to map their own communities. • Work with Data Map Wales to see how these maps can aid bilingualism at the level of geography displayed on flood zone maps, common land maps, etc.
WP.19. Aligned Welsh/English parallel text published under an appropriate licence.	We're funding Cardiff University to examine and record the databases they need to develop the cross-lingual word embeddings referred to in Work Package 22. This will involve comparing the vector scores for words in Welsh and English. Because of the differences in volumes of existing written text, Welsh might benefit from being referenced to data-rich English. The university will then be able to create a new Welsh sentiment analysis tool using systems originally developed for the English language. A further form of parallel English/Welsh text is of course translation. This is being addressed under Work Package 11 above.
WP.20. Stemmer.	We've already funded the University of South Wales to create a Welsh language Stemmer. It's available under open licence at https://hypermedia.research.southwales.ac.uk/kos/wnlt/. Over the coming months, we'll assess whether there is a need for further work to be done in this field.
WP.21. Parsers: dependency parser, constituency parser.	Developing parsers involves dependencies which we're addressing during 2020-21: e.g. word embeddings. We'll consider relevant developments in parsing when these dependencies have been addressed.

Description and *Action Plan* Work Package number	**What’s been done/What’s being done**
WP.22. Word and Term Embeddings	We’re funding Cardiff University to undertake work in this area. Word embeddings are a mathematical way of analysing the Welsh language. Each word is given a vector according to its proximity to other words. Co-relationships can then be seen when analysing the language automatically (e.g. by a bot) and confusion around the meaning of ambiguous words can be reduced as intent is more clearly detected. Welsh language vector scores will be available for downloading under an open licence by November 2020.
WP.23. Term extraction.	Cardiff University has already created a Welsh term finder called Flexi Term Cymraeg and released it under an open licence. (see https://github.com/ispasic/FlexiTermCymraeg). They did this as part of CorCenCC, which was funded by AHRC and ESRC.
WP.24. Welsh Language Named Entities	We’ve already funded the University of South Wales to create 129 lists of named entities. They’ve been released under an open-ended licence to be downloaded from: https://hypermedia.research.southwales.ac.uk/kos/wnlt/ and https://hypermedia.research.southwales.ac.uk/kos/wnlt2/ The implications of defining these Welsh language entities are that Welsh language key terms, sayings and names can be automatically tagged and interpreted as a whole concept so they are not mis-translated by computers. We’ll consider whether further work needs to be undertaken in this area in the future to increase the accuracy of machine translation, sentiment analysis, artificial intelligence etc.
WP.25. Welsh language sentiment analysis scores for broad domains, names, idioms, etc.	We’re funding Cardiff University to develop new Welsh sentiment analysis. This will enable organisations to score and highlight text from customers and clients who write in Welsh. These data can help to highlight any problems, even when there is a great deal of data to scan and analyse.
WP.26. List of Welsh Stopwords under an appropriate licence.	Stopwords are those linking words and terms—like ‘and’, ‘of’, ‘at’—which may not add much to the meaning or intent of a sentence. Being able to automatically recognise and ignore Welsh language stopwords are one of the dependencies referred to above when we discuss parsers. A list of Welsh language Stopwords created with our financial assistance can be downloaded from the University of South Wales under open licence from: https://hypermedia.research.southwales.ac.uk/kos/wnlt/

Description and *Action Plan* Work Package number	**What’s been done/What’s being done**
WP.27. Welsh language WordNet.	We funded Cardiff University to create a Welsh WordNet. A wordnet is a database of Welsh nouns, verbs, adjectives and adverbs arranged in the form of sets of synonyms and linked together by various lexical and semantic features which they share. The WordNet is available under an open licence from: https://github.com/CorCenCC/wncy. The implications of linking the meanings of Welsh words in a network like this is that it’s possible to disambiguate the meaning of certain words so that computers can better understand the user’s ultimate intent.

Glossary

List of terms used in this document, with our definitions.

Term	Definition
Acoustic model	A component of automatic speech recognition which helps with the mapping of the sounds of utterances to written letters or syllables..
Aligned parallel text	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.
Conversational AI	Some robots build things and move in the physical world. Others use natural language processing to offer appropriate and useful responses to questions and circumstances, also known as Conversational AI (artificial intelligence).
Corpus (plural: corpora)	A large collection of texts, recorded or printed. It can also be a collection of sound recordings or human gestures (i.e. sign language).
Embeddings	Machine learning algorithms that use corpora and vectors to assess a word within the context of a sentence and suggest a meaning based on probability. It’s a way of disambiguating the meaning of a word.
Hunspell	A list of words used in a spellchecker.
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 aided 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’ isn’t translated literally as the Welsh ‘smithy’ 'Dr Gof'.

Term	**Definition**
Parser	E.g. dependency parsers, constituency parsers. They 'reveal' the meaning of sentence semantics for machine translation, conversational AI etc. The computer analyses and splits a sentence and thus creates a sort of 'tree', which reveals the grammar of the sentence, citing the conceptual relation of words with each other.
Parts of speech tagger	A resource which automatically labels words in text as being 'noun', 'verb', etc.
Sentiment analysis	Facilities which enable analysis of a body of data/texts to quantify certain emotional conditions (among other things). It 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. E.g. 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', 'intolerable ' 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 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 the 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.
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 computer 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 terms within passages automatically, meaning that the terms, not the individual words, are analysed/translated.

Term	**Definition**
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 will 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 aided translation (CAT) tool.
Translation technology	Such technology can be in the form of translation memory software which inserts segments from translation memories, 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 translation, a gist translation, is required.
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.