Welsh language technology action Plan

Final Report (2018-2024)

Audience
Everyone who wants to see Cymraeg flourish in the digital world.

Overview
This report offers an overview of progress on the work packages in the our [Welsh language technology action Plan](#) since its publication in October 2018. The Plan is part of our [Cymraeg 2050: A million Welsh speakers](#) strategy (2017). Its intention was to outline technological developments so Welsh could be used in a wide range of situations, using human voice, keyboard, or other means of interaction. We published a [progress report](#) on this Plan in 2020. This report includes and builds on the 2020 report.

Steps to take
For information only.

More information
Want to know more about our work? Contact us. Here are the details:

Cymraeg 2050 Division
Welsh Government
Cathays Park
Cardiff
CF10 3NQ

email: cymraeg2050@gov.wales

[Twitter] @cymraeg

[Facebook] Facebook/Cymraeg

Related documents
[Welsh linguistic infrastructure policy (2023)](#); [Welsh language technology action Plan (2018)](#); [Welsh language technology action Plan: progress report (2020)](#); [Cymraeg 2050: A million Welsh speakers (2017)](#); [Cymraeg 2050: A million Welsh speakers: work programme 2017–21 (2017); Cymraeg 2050: A million Welsh speakers: work programme 2021–26 (2021)](#).

Contents

1 Foreword by the Minister for Education and Welsh Language

We published our [Welsh language technology action Plan](#) in 2018. The Plan aims for the Welsh language to be central to our experience of technology. Technology affects us all, and its impact is increasing. We've since published the [Digital Strategy for Wales](#), and the [Welsh language technology action Plan](#) has been an important element in implementing some of the objectives of the Strategy.

Our world changed in 2020 and much of the Welsh Government's attention was diverted to combating Covid-19. I'm grateful to our partners for their flexibility and for re-prioritising work during this difficult time. One of the things that became clear from the shift to online meetings during this period was the need for simultaneous interpretation in virtual meetings. We couldn't use our language in the same way as it had been possible to do pre-Covid-19. I'm proud to say that as part of our partnership with Microsoft, we have collaborated to create a [simultaneous interpretation facility in Microsoft Teams meetings](#). This facility is provided at no additional cost to existing Teams users. Work continues with Microsoft to further develop this facility with the hope that this will lead to the creation of similar and new resources for languages spoken around the world, based on our work here in Wales.

Artificial intelligence (AI) is now in the news almost daily. At the beginning of the Plan, we decided to prepare ourselves for its potential importance to our language. And because we prepared, we're now able to respond swiftly to new opportunities. We've also funded pioneering work at Bangor University to enable ChatGPT for their chatbot, [Macsen](#). The University is now working with the OpenAI company to improve AI provision for Cymraeg in OpenAI's most powerful system, [GPT-4](#). This follows a partnership between OpenAI and the Icelandic Government which made improvements to ChatGPT in their language. That partnership was described as a way to expand GPT-4 to new parts of the world, but also as an important step toward creating tools that could help preserve other lesser-resourced languages; the same applies to this company's work with the Welsh language.

We also fund ongoing important language technology work at Bangor University, to develop bilingual Welsh and English voices and [transcription technology](#). Computer-assisted translation was another important aspect of our Plan. We're publishing and expanding our collection of Welsh Government translation memories which are freely available for all on [BydTermCymru](#), and we want to see them shared widely. Such sharing helps ensure that the Welsh language we see is more consistent and more visible in more places.

Another important development during the lifespan of the Plan was the establishment of the Centre for Digital Public Services (CDPS). We've worked with the Centre since its inception to ensure that the Welsh language is a core part of its work, and it is heartening to see a new organisation embracing our language in such a professional way. One of the first aims of CDPS was to publish [Digital Service Standards for Wales](#) which emphasise the importance of developing services in both Welsh and English. A big step for this work was the launch of [Trio Writing](#) in collaboration with us at the 2023 Llŷn ac Eifionydd National Eisteddfod. This booklet, like all of CDPS's and our

work in terms of Cymraeg 2050, is people-focused. This approach needs to be at the heart of our technology development work too. Language is a people thing, and technology needs to help people use Welsh. We don't do technology for technology's sake.

A core principle of the Plan was our emphasis on developing resources and components under a suitably permissive licence so more developers and Welsh speakers could access them free of charge. That philosophy continues. Our job at the Welsh Government is to drive the developments, and then encourage developers to use those components widely. I'm very proud of the new resources which are now available to use and reuse as a result of our work. You can read about them in the following sections.

This was an ambitious Plan—one that required collaboration to achieve the objectives. The resources we created as part of this Plan go a long way towards ensuring the robust infrastructure that we need to realise our vision for our language, but language technology and artificial intelligence are still racing ahead. There's still a lot of work to be done. We need to ensure that new systems help people live and work bilingually. We have invested in the technical infrastructure of the Welsh language, and the resources now need to be shared and maintained for future generations. Technology is constantly evolving, and the Welsh language must evolve with it.

Jeremy Miles MS

Minister for Education and Welsh Language

2 Glossary

Definitions of the terms we use in this report.

English term	Welsh term	Definition
acoustic model	model acwstig	A tool that helps a speech recognition system know which word is pronounced according to the sound of each word or syllable.
augmentative and alternative communication (AAC)	cyfathrebu estynedig ac amgen (AAC)	A variety of strategies and tools to help people who have difficulty with speech. It may use simple letters or pictures, or sophisticated computer systems. AAC helps someone communicate as effectively as possible, in as many situations as possible.
chatbot	sgwrsfot	A 'chat' machine with a conversational artificial intelligence capability. A person asks a question and receives a response from a chatbot.
conversational artificial intelligence	deallusrwydd artiffisial sgwrsiol	A computer application that processes natural language to understand questions and offer appropriate and useful responses.
corpus (corpora)	corpws (corpora)	A large collection of recorded or printed texts. A corpus may also be a collection of audio recordings or human gestures (ie sign language).
defaulting	rhagosod	Options pre-selected without a user having to do anything.
end users	defnyddwyr	People who use a computer or technical system.
generative artificial intelligence	deallusrwydd artiffisial cynhyrchiol	Technology that allows computers to create new content, such as text, images, sound, music, code, and even scientific data. Unlike traditional artificial intelligence (AI) methods that focus on analysing and interpreting existing data, generative AI aims to learn the underlying patterns and structures of a dataset and then generate entirely new content based on that information.
machine translation	cyfieithu peirianyddol	A service, eg Google Translate/Microsoft Translator (among others) that automatically translates text from one language to another. Translation memories (see below) can be used to train or improve systems such as these.
named entities	enwau endidau	Words and terms that should be treated as one specific concept by computer systems, eg names or places.
neural networks	rhwydweithiau niwral	A way of processing data that mimics a human brain. Connections are created and modified between artificial 'neurons' during the training

English term	Welsh term	Definition
		process. A lot of data is needed to develop a neural network.
noun chunker	talpiwr enwau	Software that groups words together to form one nominal concept, eg 'little red car.'
page views	edrychiadau	In the context of website and social media user statistics, one example of a page (or resource, video etc) being opened.
parallel text	testun cyfochrog	Aligned texts, which is the process of arranging English and Welsh text eg, into corresponding parallel segments.
parser	parsydd	Software that reveals grammatical syntax in a sentence for machine translation, conversational artificial intelligence, etc.
part of speech tagger	tagiwr rhannau ymadrodd	A tool that automatically labels words in sentences, eg 'noun', 'verb'.
sentiment analysis	dadansoddi sentiment	A process to automatically analyse a body of data/texts to quantify specific emotional conditions. It can use a list of words or terms with an equivalent score to assess and score text in various domains, eg health, broadcasting.
spaCy	spaCy	'Python' computer language library for natural language processing.
speech-to-text	lleferydd-i-destun	Software that turns spoken language sounds into written text.
stemmer	boniwr	A software or script that shortens the end of words to reveal the root of the word, eg 'developed' to 'develop-'.
stopwords	ataleiriau	Functional words such as 'the', 'or' that don't add to the themes or meaning of a piece of text. Technologists use a list of stopwords to filter the text and leave keywords behind.
term extraction	echdynnu termau	A computer script that studies the frequency of word order and creates sets of words that are likely to be terms. 'Datgelydd termau' is also used in Welsh.
text-to-speech	testun-i-leferydd	Software that turns written text into spoken language sounds.
touch type	teipio cyffwrdd	Typing using each finger and without looking at the keyboard.
training data	data hyfforddi	Training data—in the form of text, audio, etc—which feeds into language technology developments.
Translation memory	cof cyfieithu	Sentences or segments in one language aligned with a corresponding parallel translation in another language.
voice banking	bancio lleisiau	The process of storing recordings of a person reading a script or speaking naturally. One use

English term	Welsh term	Definition
		is to use software to break up the recording into a series of pieces to create a custom synthetic voice.
Wikidata	Wikidata	Collection of labels of topics covered within Wikipedia articles.
word and term embeddings	mewnblaniadau geiriau a thermau	Machine learning algorithms that use corpora and vectors to consider the word within its context and offer meaning based on probabilities and previous data.

3 Introduction

This retrospective report examines what we've achieved in technology and the Welsh language since we published our [Welsh language technology action Plan](#) in 2018. The report focuses on Welsh Government-funded work but, for the document to show the full range of what's available, it also includes work by others whilst—of course—recognising and acknowledging their contribution. Diolch yn fawr, therefore to the research councils, to Universities, and to the many volunteers who give generously of their time to the world of technology and Welsh language.

Here are some highlights of the Plan:

- ChatGPT can be used within Bangor University's chatbot [Macsen](#), and the University has started working with the OpenAI company on Welsh language models for [GPT-4](#) to improve the output for the language.
- Bangor University released bilingual text-to-speech voices, their [Transcriber](#), and further improvements to their chatbot, [Macsen](#). These improvements in turn fed into their voice banking project, [Lleisiwr](#).
- We made [Cysgliad](#) available free of charge to individuals, organisations with fewer than ten members of staff, and to the whole education and third sectors.
- We defaulted the language of the Office365 interface to Welsh in Welsh medium schools, reducing 'friction' for the users and making Welsh easier to use.
- The Welsh-language [Wikipedia](#) (Wicipedia) now has more contributions, with the number of new articles increasing since the Plan was published. The content created by the volunteers adds to the Welsh language presence on the internet, which helps to increase social capital and encourages more writing in Welsh.
- We're freely sharing a collection of Welsh Government translation memories on [BydTermCymru](#) and we've funded work to create health and care and legislation 'domain specific' machine translation systems too.
- As part of our partnership with Microsoft, we have collaborated to create a [simultaneous interpretation facility in Microsoft Teams meetings](#). This facility is provided at no additional cost to existing Teams users. Work continues with Microsoft to further develop this facility with the hope that this will lead to the creation of similar and new resources for languages spoken around the world, based on our work here in Wales.

You can read full details of all the work done in the following sections.

4 Background

The priorities of our [Welsh language technology action Plan](#) (2018) were to develop and share new infrastructure for our language, specifically in:

- Welsh language speech technology
- Computer-assisted translation
- Conversational Artificial Intelligence

One of its main principles was to develop a culture of open innovation. We wanted more digital resources and Welsh language data to be available, without unnecessary conditions limiting their use, and shared under a suitable licence—open and permissive wherever possible—to support efforts to innovate digitally for the benefit of our language.

We've long recognised the need to strengthen the infrastructure for Cymraeg. This is central to our strategy, [Cymraeg 2050: A million Welsh speakers](#). We want to ensure that we can use Welsh in as many situations as possible—to contribute directly to our goal of doubling the numbers of us who use our Welsh every day.

During this Plan, the way we work changed completely due to Covid-19. We had to move swiftly to working online and it became clear that we all needed to be flexible and creative in working together in the then 'new normal'. The National Centre for Learning Welsh, for example, accelerated the delivery of online provision so learners could continue to learn online. As part of our partnership with Microsoft, we have collaborated to create a [simultaneous interpretation facility in Microsoft Teams meetings](#). This facility is provided at no additional cost to existing Teams users. Making Welsh easier to use—that's our goal when working with big companies; that and taking the 'friction' out of using Welsh for end users.

5 Work Packages

Below, we detail the original work packages of the 2018 Plan, with an update (as of February 2024) for each work package. The 'Glossary' section at the beginning of the document may help you with any unfamiliar terms.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018

1	Welsh language speech-to-text facilities and components.	Software created under an appropriate licence, work undertaken to convince relevant organisations/corporations to adopt that software.	Speech recognition has become more important in recent years, not only as an interface with AI but also as an automatic way to generate subtitles and transcripts of videos and meetings.
			
			We've funded Bangor University to release a Transcriber that can automatically generate subtitles to Welsh-language videos. Recent developments and improvements reduce the amount of post-editing needed on this system.
			
			As a result of developing the Transcriber, the University has been able to release the following additional resources under a permissive licence:
			
			• Corpus of training sentence prompts for Welsh language speech recognition.
			• Language normalization component for text-to-speech (this corrects things like "14:30" to "half past two" instead of "one four colon three zero"). It is operational in the following bilingual voices:
			    o northern female
			    o southern female

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			- northern male - southern male - A bilingual acoustic model that can recognise both a Welsh and English voice within the same sentence. This may lead to facilities that could transcribe bilingual meetings, for example. Many of the speech recognition models available around the world are monolingual models. Ie, they recognise one language at a time. In places where two languages are spoken, where both of those languages are used daily by the same people, and where there are bilingual meetings and situations, we need a model that recognises two languages at the same time. This is complex and innovative work that's not widely available in the world of speech recognition and therefore in its infancy in terms of Welsh. - A Welsh punctuation and capitalisation model to improve the Transcriber by placing the full-stop in the correct places. - A dictation language model that ensures the sentences in the output are understandable. - A new speech model based on unscripted speech data. The original Plan also encouraged and facilitated the use of previous crowdsourcing work, where possible, to further increase the amount of training data available. Some examples of this are the crowdsourcing speech-capture app Paldaruo (Bangor University) and the Mozilla campaign Common Voice (work

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			package 3). Data like these help improve the models listed above to create a more accurate Welsh output. We've also funded 16 augmentative and alternative voices [AAC] for children and young people who rely on technology to communicate. You can read about this below in work package 9.
2	Welsh language machine learning and conversational artificial intelligence.	Software under an appropriate license.	We've funded Bangor University to further develop a virtual assistant Macsen (now available on iOS and Android). It uses audio corpora and transcripts of the work that from work package 3. Also, ChatGPT now works in Welsh within Macsen as well and gives promising results. This allows Macsen to be asked more complex questions, after which he can then produce more creative responses. Generative AI and ChatGPT are becoming increasingly important in language technology, and this is a major step in future-proofing the Welsh language. With the help of our grant, Bangor University is now working with OpenAI to refine the models in OpenAI's most powerful chatbot, GPT-4. The University's developing a set of evaluations for GPT-4's performance with Welsh to scrutinise the model's performance in the context of bilingual Welsh names, Welsh language yes/no questions and other syntactic and grammatical issues, and to develop ways to improve its output.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			Macsen also includes an intent parser (this is a tool that senses whether someone is asking for the news or the weather, for example). These skills are quite different from ChatGPT and more like other smart assistants in our homes, for example. Macsen therefore knows when to send a question to ChatGPT and when to treat it internally as a request to turn off the lights through Phillips Hue, for example.

Since we started conversations with the Amazon Alexa team in Seattle, Welsh has now been included as one of 51 languages in the AI database Amazon: MASSIVE. MASSIVE is a parallel database of up to a million annotated phrases across 51 languages for natural language processing and having Welsh built into this is a step further in teaching Alexa to speak the language.

We're also working with the Government of Catalonia on their AINA project. This is a project between the Government of Catalonia and the Barcelona Supercomputing Centre to develop several language technology packages relating to translation, speech and AI and we've been able to provide data for their Welsh language model. |
| 3 | Welsh language audio corpora for annotation, to train speech-to-text and other technologies. | A collection of transcribed audio recordings suitable to train systems. | By collecting more Welsh audio data and annotating it for training, Welsh language speech technology will improve.

So we've continued to fund Paldaruo—a crowdsourcing app to collect Welsh language speech data, and we've funded work to |

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			increase the number of Welsh speakers contributing their voices too.

In addition, Bangor University worked with Mozilla to contribute Paldaruo data to its Common Voice crowdsourcing campaign. Common Voice collects speech data in languages around the world to increase the volume of data and improve speech technology, especially for lower-resourced languages. The Welsh corpus can now be downloaded and used under a permissive licence. At the time this report was drafted, 156 hours had been recorded in Welsh.[1]

It's worth noting that similar data collection campaigns are being undertaken for lower-resourced languages around the world and that the Welsh Government also contributes to them where possible. Through our work with the Adapt Centre at Dublin City University for example, we contributed our voices to the WebNLG Challenge 2023. The purpose of the task was to turn engineering data into everyday speech which is more useful for training (the results are now available online).

Bangor University has also created openly available verbatim transcripts for 30 hours of speech data from podcasts and other sources. The corresponding audio data has been distributed under |

---

[1] https://commonvoice.mozilla.org/cy/datasets (accessed 02/02/24)

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			an open license so others can train speech recognition models with the data.
			The translation company Cymen has won an 'Arfor' grant to follow the example of languages with larger volumes of data, to pay Welsh speakers to contribute suitable speech data—ie audio files of people speaking Welsh with a text transcription. The data will be openly available so it will help improve the standard of Welsh language transcription, and will be available to other projects, for example Common Voice (above).
			We've also funded PhD research at Bangor University in partnership with KESSII. This research uses Paldaruo and Common Voice data, for example, to improve our understanding of how computer systems process the Welsh language.
			The nature of collecting data in this way isn't to amass finished resources but a process of creating, contributing to, and maintaining components over the long term as all languages constantly evolve. This is a challenge for many other lesser resourced languages around the world.
4	Improving user experience of Welsh language technology through behavioural economic techniques.	Research Reports, practical guides, techniques to increase the number of the Welsh language locale users (be they individual users or	Work packages 7 (defaulting interfaces to Welsh) and 13 (automation of Welsh language services) also apply to this work package as they add to the user experience when using online services.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
		institutional users), indicators of users that use Welsh services/interfaces.	This work package focuses on improving user experience using behavioural economics techniques. One of these techniques is 'EAST' (easy, attractive, social, topical). For example, it's possible to increase the use of Welsh by making it easier to choose online (ie reducing behavioural 'friction'). While keeping a focus on 'EAST' features, ie on the needs of the users, the experience of using Welsh in technology is more accessible, more in keeping with 'EAST'. Since the Centre for Digital Public Services (CDPS) was established, we've worked together to ensure the Welsh language is at the heart of the new organisation. One of their first objectives was to publish the Welsh for Digital Service Standards for Wales which emphasise the importance of developing services in both Welsh and English. During the 2022 Ceredigion National Eisteddfod, we worked with CDPS and others to carry out research with visitors to understand what would help Welsh speakers to use more services in Welsh. The results were discussed at a public session with the Minister for Education and Welsh Language. This work fed into our collaboration at the 2023 Llŷn and Eifionydd National Eisteddfod, where we launched the book Trio Writing in collaboration with CDPS. Trio writing is a technique of creating content in a way that reflects English and Welsh while keeping the content's meaning or service at its centre. It's a way to make both versions easy to use, again making it more 'EAST'.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			On our Helo Blod website there's a list of Welsh and bilingual technology resources. We've also developed and published a bilingual technology toolkit for good user experience on the Helo Blod website to help companies and developers design digital services that work bilingually and are easy to use in both languages. We've incorporated this guide into our procurement processes and report on this work below.
5	Frameworks to personalise text-to-speech and bank individual voices.	Software and instructions available under appropriate licence.	Work packages 1 (speech-to-text) and 2 (artificial intelligence) also apply to this package.

We laid the foundations of Welsh language voice banking by funding Lleisiwr, created by Bangor University. As part of Lleisiwr, individuals can record their voices by reading a script, which Lleisiwr turns into a synthetic version of their own personal voice. This can be used for individuals who may subsequently lose their voices for medical reasons, for example. We've funded the University to carry out further work which includes the latest text-to-speech enhancements, and Lleisiwr 2 is now available. It's used in the health sector to help Welsh-speaking patients use both Welsh and English after their treatment. The service can be accessed through a speech therapist or medical specialist.

The University is also working on a synthetic voice cloning system to tailor the synthetic voices to the individual even more. This enables the same voice to be used for English and Welsh which improves the bilingual user experience. |

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			In addition to individualistic voices, it became clear in discussion with stakeholders that generic synthetic voices needed to be developed so anyone could use them. So we funded Bangor University to develop bilingual voices (listed in work package 1, speech-to-text) that would be able to cope with codeswitching between Welsh and English. This built on Gwyneth and Geraint's voices, we and the former Welsh Language Board funded prior to the 2018 Plan. As a result of text-to-speech development and voice banking, these other components are also available to use: • A collection of recording prompts for voice talent • A normalization component for numbers, abbreviations and time, as noted above.
6	Interactive content and software for Welsh learners.	New digital resources available to help people learn Welsh and to improve oral and written Welsh.	The National Centre for Learning Welsh accelerated the roll-out of online Welsh learning provision in response to Covid-19 in 2020. The sector used technology to facilitate its work through the pandemic and to offer online lessons. Learners were therefore able to continue to learn Welsh through lockdowns, as well as keeping in regular contact with other learners to practise their Welsh. These developments, and the training given to tutors during the pandemic, have enabled the Centre to widen its offer to learners. By 2023, the Centre offered a choice of in-person, online, or hybrid courses.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			During the pandemic the Centre also hosted 'Cymraeg yn y cartref' sessions on Facebook, providing opportunities for families to learn and practise their Welsh on a daily basis when schools were closed. The Centre has also increased the offering of resources available on its website. They've launched 'Gwiriwr Lefel', a tool that analyses the user's proficiency in Welsh. They've also launched more 'Cymraeg Gwaith' self-study resources for workers in specific sectors, for example the education workforce, care sector workers, as well as those working in early years and the tourism sector. They've also developed dedicated digital resources to teach Welsh to young people between 16-25, and digital resources are also available to enable people to learn Welsh through other languages. Since 2020 the Centre has also forged partnerships with Say Something in Welsh (SSiW) and Duolingo. The Centre provides a free SSiW resource for all its learners, and the Centre and SSiW have worked together to create digital resources to support young people to learn Welsh, and to create resources to learn Welsh through other languages as noted above. Between November 2021 and November 2023 the Centre looked after the content of Duolingo Welsh courses. In that time the Centre ensured that the Welsh Duolingo course was aligned with the Centre's 'Sylfaen' and 'Mynediad' courses so the Centre's learners could use Duolingo as a resource to review and practise their Welsh.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			With the help of a Welsh Government grant, Bangor University has also updated Vocab. This is a resource highlighting key Welsh words and showing an English translation above. Websites of all types can use Vocab and the University is currently testing the latest version, which includes text-to-speech. We've also worked closely with the CorCenCC project, which has produced a number of resources to help people learn Welsh. For example, the Tiwtiadur which helps teachers set tasks for their learners. As resources created under this Plan are released under an open licence, volunteers have been able to actively use them for their own purposes, and we're very grateful to them for giving so freely of their time and expertise. CorCenCC resources, for example, are being used within hir-iaith, which is a resource to check grammar and analyse Welsh text. Using CorCenCC's part of speech tagger, it's possible to disambiguate the meaning of words, analyse sentiment and ensure grammatically correct text. All of this helps learners understand what individual words mean, their grammatical purpose in sentences, their different forms, conjunctions, and the reason for these variations.
7	**Education and Skills**	Lessons learned logs, case studies, a report on the	We noted in our previous progress report the work we'd done on this. We successfully defaulted Microsoft Office365 Hwb interfaces to Welsh (with 78,086 users). In so doing, all learners in Welsh-

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
	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.	Welsh language UI in a sample of organisations.	medium schools continue to use the resource in Welsh without having to do anything to opt in themselves. In reducing the friction of language choice, the bilingual user's experience of the Welsh language is easier and more accessible, again drawing on the principles of 'EAST' behavioural techniques (as discussed in work package 4). We're also working on other avenues of 'defrictionising' language interface choice. For example, we have worked with Stable to make it more 'EAST' to use the Microsoft Welsh language interfaces on Windows and Microsoft applications. Stable has created a new initiative called CALI (Customised Automated Language Interface)—a process designed to automatically change a user's entire interface to the Welsh language. We're currently testing this functionality and we believe it will be used, in due course, by many organisations across Wales. We've also worked with Google to ensure that Google for Education is available in Welsh via Hwb. The same goes for and Adobe Spark. This collaboration has created a fairer and more inclusive learning experience for learners in Wales and enables Welsh-speakers all over the world to access Google Workspace and ChromeOS in their chosen language too.
8	Promote Welsh language technology	Relevant resources available, training carried out.	We've worked with partners to support the development of Welsh language coding learning resources.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
	and coding resources to teachers and children and others.		We've worked with the BBC to promote their coding campaign to primary schools. This includes new learning resources on the Hwb website. Suggestions for teachers are part of the resources, which map to the relevant parts of the curriculum. In November 2023, Technocamps presented a 'Hwb Presents...' introductory webinar on delivering the Curriculum for Wales using micro:bits. The webinar and resources were promoted via Dysg, Hwb and Technocamps. In-person events are planned in 2024 as the micro:bit devices continue to be rolled out to schools across Wales. Technocamps have also been active at the Urdd Eisteddfod highlighting what they can offer as support to Welsh teachers and schools, highlighting the link between problem solving while playing games and computer mindset skills.
9	Create and/or develop facilities to assist Welsh speakers with additional learning and/or accessibility requirements.	Relevant resources available.	Work packages 1 (speech-to-text) and 3 (voicebanking) also apply here. We've funded 16 augmentative and alternative communication [AAC] bilingual young voices for children and young people who rely on technology to communicate. These include voices with northern and southern Welsh accents for young and teenage boys and girls. These voices will be available through the National Centre for Electronic Assistive Technology which works with professionals across Wales to provide tools to improve the quality

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			of life and independence of people with conditions that affect their access to everyday activities, including communication. We've also commissioned a resource called 'Teipio Testun' (available through a Hwb account). This is an accessible activity on the Hwb website to support Welsh and English speakers at school in learning to type touch. Discussions with additional learning needs (ALN) experts noted this type of resource as something that would help teachers who teach blind or visually impaired learners, for example.
10	**The Workplace** Ensuring suitable English/Welsh and Welsh/English machine translation systems for different linguistic domains and registers.	Relevant training data available under an appropriate licence; Application Programming Interface to Welsh/English and English/Welsh translation engines available.	We've funded Bangor University to create domain-specific machine translation engines for Health and Care and Legislation. When focusing on a specific domain, results tend to improve compared to more general machine translation engines. Bangor University has now made the translation engines available within Wordfast, a popular professional translation memory software package, and they are currently being tested by the University's Translation Unit. The University has also extended its work from work package 2 (artificial intelligence) and refined OpenAI's GPT 3.5 with parallel data to obtain promising results when GPT 3.5 is used as a translation engine. Bangor University's chatbot, Macsen (see work package 2) also now includes a voice-to-voice translation feature that uses a machine translation model on top of speech and text-to-speech recognition.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			We've funded Cardiff University to use cross-lingual embeddings (words with similar meanings) so the Welsh language can benefit from languages with a greater volume of language data, such as English, to improve output in Welsh. This helps to establish principles that will be used in other aspects of natural language processing in Welsh. We also freely share translation memories under an open licence which can be used as parallel bilingual data for natural language processing in Welsh. The data generated from this can be fed into automatic translation engines (see work package 11).
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	Shared computer-assisted translation networks; data on the use of technology and the translations thereby undertaken.	Our Translation Service has shared 144 translation memories thus far on the BydTermCymru website. Any translators may upload these memories to their own translation memory systems and use them for free. In addition to this, the resources on BydTermCymru include a complete version of the TermCymru terminology and titles database which can be downloaded and used within any translation memory or terminology management system. We've explored how to facilitate the sharing of translation resources between different organisations. Some have raised concerns about sharing confidential and sensitive data that may be in the memory. So, we're currently researching 'scraping', tagging

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
	share translations in real time.		and aligning parallel Welsh/English public texts, and looking at how to use the data to create translation memories and new domain-specific translation engines (see work package 10). As this data is already public, it should alleviate concerns about confidentiality. Bangor University has created the Open Translation Memories resource, which is a way to search for a word in translation memory databases.
12	Modify, where relevant, procurement processes, so the Welsh language is a consideration in technology from the outset.	Details on the Welsh Government's website that explains the amended procurement processes with attendant guidance.	We've designed and published a bilingual technology toolkit for good user experience on our Helo Blod website (a website for business advice on how to use more Welsh). The toolkit is there as a resource to help anyone or any organisation to create bilingual digital services that offer a smooth and easy experience for their users. The toolkit can also be included in tender specifications and/or various other proposals that require technology. It leads the reader through a process to ensure the product offers a smooth bilingual service. It also signposts to more detailed technical advice if necessary. When used in conjunction with Trio Writing (see work package 4), this can help you create easy to find, easy to use, and easy to read Welsh language services. The toolkit also now appears on the Cyd website—a specialist support website for procurement staff. As one of their resources,

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			the toolkit will appear on the 'procurement journey' as a specific step that needs to be considered in tender specifications. The toolkit will also be incorporated into the statutory guidance for the Social Partnership and Public Procurement (Wales) Act. The Act aims to improve public service delivery and well-being in Wales, so this is another way of ensuring that the needs of the bilingual user are considered from the outset in the future. Once again, the principles of this work will apply to many multilingual organisations around the world, and we share it through our national and international partnerships.
13	Explore the potential of technology to facilitate and/or automate Welsh language services eg automatically redirect phone calls to Welsh speakers within organisations.	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.	As part of the our policy More than just words: Welsh language plan in health and social care, data has been collected to bring together the available sources on the Welsh language skills of workers in health and social care. The data show that situations often arise, in GP practices for example, where users aren't even aware that it is possible to choose a different language from English. We're now working with private companies to explore other ways to remove the friction from language choice. This follows the success of defaulting the Microsoft 365 interface on Hwb to Welsh for Welsh medium Schools (see work package 7). This again draws on the principles of 'EAST' (see work package 4 on behavioural economics), and also feeds into our strategy for

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			internal use of Welsh: Cymraeg. It belongs to us all (specifically, Action 8: Increase use of Welsh language ICT resources and develop a friction-free experience for users).
14	A list of Welsh language and bilingual ICT resources available in the workplace, also noting gaps in provision.	A list of available and missing resources.	We've published a Welsh language technology resource list on our Helo Blod website. It offers support in developing new software, resources, or services in Welsh.

You can also use the Helo Blod website to make enquiries about Welsh language software (and its use in business). It's fast, it's friendly and it's free. |
| 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. | Reports on the number of workshops and the resultant outputs. | We've funded work at the National Library of Wales which has created a community of volunteers to create original Welsh-language content on Wici platforms.

Between 2018 and 2023,
- The number of new Welsh articles on Wicipedia increased from 101,008 to 279,643
- The number of Wikidata items with Welsh labels increased by 1,278,788 to 3,593,389
- Wicipedia page views increased from 51 million to 108 million between 2018 and 2023[1] |

---

[1] https://stats.wikimedia.org/#/cy.wikipedia.org/content/pages-to-date/normal%7Ctable%7C2018-07-27~2023-09-03%7Cpage_type~content%7Cmonthly (accessed 02/02/24)

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			We're extremely grateful to the volunteers who author and edit Wicipedia for giving so freely of their time to contribute to this work.

A big diolch too on our part to the volunteers of the Meddal.com website too, who translate software that may be of interest to the Welsh language community and promotes software through social media and the website itself.

Work package 18 discusses the work Mapio.Cymru has been doing with our funding, including running public workshops to encourage the public to add to the Welsh language interactive OpenStreetmap.

Work package 8 introduces Technocamps, which has been running public engagement activities at the National Eisteddfod, for example.

Another example of our work in content creation is Hac y Gymraeg (the Welsh language Hacathon which we've funded M-Sparc to organise).

One of the first Hac y Gymraeg winners, Mwydro worked with north Wales secondary and primary learners to design and promote a series of Welsh place name GIFs stickers for social media. The rationale for the project's success wasn't necessarily the GIFs it created. But rather, it was the fact that they were created in Welsh, by bringing participants together who may not have used their |

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			Welsh socially before and therefore experience Cymraeg in a new, out-of-classroom way. Another Hac y Gymraeg winner was DarllenCo, a subscription system to digitally access original children's books where readers can hear a specific word and learn more about what it means, and where they can also collect 'streak' points and badges by reading; and Si-Lwli, a Welsh-speaking toy company: - Our funding to DarllenCo's via Hac y Gymraeg was key to their journey to launch their platform for schools (in November 2022). The funding allowed them to create more content and to add more functionality. Just under 40,000 people now benefit from the platform. With M-SParc's support, DarllenCo has also been able to reach the market and deliver a sustainable business based on the Welsh language. - Hac y Gymraeg also funded, Si-Lwli the first Welsh dolls of their kind, 'Deian' and 'Loli'. These dolls are a new way of bringing the Welsh language to young families. We've also funded conferences and symposia at Bangor University that draw together experts in language technology from all over the world, the most recent being a symposium in December 2023. OpenAI made the keynote presentation on how to improve support for lower-resourced languages in the field of generative AI.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
16	Long-term support for the development of the linguistic infrastructure of the Welsh language, including corpora, lexicographical and terminological resources.	Downloadable resources released under an appropriate licence.	We published our Welsh linguistic infrastructure policy in June 2023. This policy outlines what we'll do to develop all elements of linguistic infrastructure, including dictionaries, terminology and corpora. There are already several projects and resources available in this context: dictionaries, terminology resources, corpora and standardisation work, but there is a lack of co-ordination between them. We've established a new Unit within the Welsh Government to make sure this co-ordination happens. The Unit will also promote these resources through a new website. We've set up a Welsh Language Standardisation Panel to solve orthographic problems. The panel includes academics, lexicographers, linguists and translators who are experts in their fields. We'll share the panel's decisions with relevant stakeholders and organisations, with the aim of seeing resources updated to reflect those decisions. There's also now a new section on BydTermCymru called Subject-specific Lists. It contains three new aspects: • Glossaries, which include lists of terms in specific fields, including race and ethnicity terms, LGBTQ+ terms, and Covid-19 key terms • Place-names in Wales, which includes lists of the names of all electoral wards in Wales and the names of all monuments in Cadw's care, as well as information about

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			the List of Standardised Welsh Place-names by the Welsh Language Commissioner. • International place-names, which include lists of the names of states, UK Crown Dependencies and UK Overseas Territories
17	Welsh spellcheckers, grammar and mutation checkers available free of charge.	Relevant software available under an appropriate license.	As a result of our grant to Bangor University, since May 2020 Cysgliad has been available free of charge to individuals, all schools in Wales, and organisations with fewer than 10 employees. Our initial intention in releasing Cysgliad was that it would specifically help learners to create content and to do schoolwork in Welsh at home during lockdown. Within two weeks of the announcement that the software was available for free download, 5,032 copies had been downloaded. At the time this report was drafted, 12,586 had been downloaded. With our funding, Bangor University has also developed the following linguistic components: • Update to the Lemmatizer (which identifies all forms of a word) • Welsh and bilingual part of speech tagger (software that reads text in a language and designates a part of speech to each word) • New versions of Hunspell word lists (Hunspell CY and Hunspell CY Verbal), a comprehensive list of Welsh words

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			that can be used to highlight where words are incorrect. Hunspell is therefore used as a Welsh spell checker. • An open lexicon of Welsh word forms and their parts of speech and morphological information
18	Interactive maps with Welsh language versions of place names that can be embedded within web pages.	Relevant cartographic resources with Welsh language place names.	We've funded Data Orchard CIC to enhance and expand the reach of the Welsh language interactive map Mapio Cymru. They're also working with OpenStreetMap volunteers in Wales to make the map more comprehensive. A Welsh language map can be embedded on a website, for example: • Mentrau Iaith Cymru have used the map to show the location of a gig on a Welsh map on their own website and so has Y Cymro website. • Since June 2023 MySociety use the Welsh language map in their Welsh language service on FixMyStreet. • MapDataCymru is used by several other public organisations, eg Transport for Wales, Natural Resources Wales and local authorities. This enables specific features to be mapped on top of Welsh mapping layers provided by MapioCymru.
19	Aligned Welsh/English parallel text published under an appropriate licence.	Texts available under an appropriate licence.	Collecting and releasing translation memories under a permissive license is the main element of this work package. At the time this report was drafted, 144 translation memories were available to download from our BydTermCymru website. Bangor University has also created the resource Open Translation Memories, which is a

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			way to search for a word in a translation memory database as well (see also work package 11). We've also funded [Cardiff University] to undertake cross-lingual development to understand those situations where Welsh can benefit from the larger data available in English (see similar work in work package 22 on embeddings). By using a part of speech tagger, a Welsh language system can benefit from work previously done for English.
20	Stemmer.	Software available under an appropriate licence.	In the lead up to the Plan, we funded the University of South Wales to create a Welsh language [stemmer] that reveals the stem of a Welsh word (an example in English would be eg 'developed' to 'develop-'). Another Welsh language [stemmer] is available from the CorCenCC team.
21	Parsers: dependency parser, constituency parser.	Software available under an appropriate licence.	We've funded Bangor University to develop a [dependency parser] that reveals the grammatical syntax of sentences. It gives more context to a sentence or question to help a chatbot, for example, respond better. Work on the constituency parser didn't continue as originally envisaged because that technology is no longer as relevant now that artificial intelligence uses different techniques.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
22	Word and Term Embeddings.	Embeddings available for Integration into translation engines, speech-to-text technology, translation automation systems etc.	We've funded Cardiff University to carry out work on embeddings, which arrange words according to their likelihood/relationship with each other. This is underway in Cardiff University's experimental interactive Thesaurus. Bangor University has also carried out work on this as part of their wider natural language processing work.
23	Term Extraction.	Software available under appropriate license.	During the term of this Plan (but funded by Cardiff University), Cardiff University created the FlexiTermCymraeg term detector and released it under a permissive licence. It combs through texts to highlight a word, or groups of words, that could be terms. A similar resource is also available as part of commercial companies' translation software. In some cases, named entities are counted as terms. We've funded Bangor University to release a Welsh-language model for spaCy that identifies and highlights named entities, among other skills.
24	Welsh Language Named Entities.	Relevant lists available under an appropriate licence.	Prior to this Plan, we funded the University of South Wales to create a list of 129 named entities that have been published under a permissive licence.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
			We've funded Bangor University to produce: • Entity Identifier, a component for identifying entities such as names of people, places and organisations within the text • A Welsh noun chunker for the word processing library spaCy (treating 'little red car' as one entity instead of three) • English named entity recognition model trained on Welsh entities (for example for a bilingual sentence)
25	Welsh language sentiment analysis scores for broad domains, names, idioms etc.	Sets of scores available under appropriate licence.	We've funded Cardiff University to work on Welsh sentiment analysis, which is a computerised way of analysing a sentence and identifying whether it conveys a positive or negative attitude. We funded the University of South Wales used WLNT2 tools to score tweets' sentiment prior to this Plan.
26	List of Welsh Stopwords under an appropriate licence.	Relevant lists available under appropriate licence.	We've funded Bangor University and the University of South Wales to work on this. A list of 488 stopwords is available from Bangor University. It weeds out functioning words ('a, the, the') to leave keywords behind. The University of South Wales also has a list of stopwords. Bangor University's Welsh language stoplist is more comprehensive and is designed to roughly match the English equivalent on spaCy. The intention is that by using them side by side it creates consistency between the two languages.

Work package Number	What (wording from 2018 Welsh language technology action plan)	Resource wording from 2018 Welsh language technology action plan)	What we've achieved since 2018
27	Welsh language WordNet.	Welsh WordNet available under an appropriate licence.	Work had been started on this at the beginning of the Plan at Cardiff University, with our funding. Word embeddings have led to better ways of connecting words, terms and entities. Cardiff University has worked in partnership with Lancaster University on a new Welsh Thesaurus which benefits from this (see work package 22).

6 Links

Here is a non-exhaustive list of relevant links. Please note that not all have been funded by the Welsh Government.

Resource	Link
AINA	https://smartcatalonia.gencat.cat/en/projectes/tecnologies/detalls/article/AINA
Amazon: MASSIVE	https://www.amazon.science/code-and-datasets/massive
Bangor University Welsh Lexicon	https://github.com/techiaith/lecsicon-cymraeg-bangor
Bangor University stopwords	https://github.com/techiaith/ataleiriau
Bilingual acoustic model	https://huggingface.co/techiaith/whisper-medium-ft-cy
Bilingual technology toolkit for good user experience	https://www.gov.wales/bilingual-technology-toolkit-good-user-experience
BydTermCymru	https://www.gov.wales/bydtermcymru
BydTermCymru Glossaries	https://www.llyw.cymru/bydtermcymru/geirfau
BydTermCymru international place-names	https://www.llyw.cymru/bydtermcymru/enwau-lleoedd-tramor
BydTermCymru place names in Wales	https://www.llyw.cymru/bydtermcymru/enwau-lleoedd-yng-nghymru
BydTermCymru Translation Memories	https://www.gov.wales/bydtermcymru/translation-memories
Common Voice Cymraeg	https://commonvoice.mozilla.org/cy
CorCenCC	https://corpus.corcencc.org/?language=en
Cyd	https://cyd.cymru/
Cysgliad	https://www.cysgliad.com/en/
DarllenCo.	https://www.darllenco.wales/
Dependencies parser	https://github.com/techiaith/parsiwr-dibyniaethau/releases/tag/23.03
Dictated language model	https://github.com/techiaith/docker-huggingface-stt-cy/releases/tag/22.01
Digigrid	https://digigrid.cymru/
Domain specific machine translator	https://cyfieithu.techiaith.cymru/?lang=en
Embeddings	https://github.com/techiaith/word2vec-cy-demutated

Resource	Link
	https://github.com/cardiffnlp/en-cy-bilingual-embeddings
English named entitiies recognition model	https://github.com/techiaith/spacy-wales-en-ner-model
Entity Identifier	https://github.com/techiaith/spacy_cy_tag_lem_ner_lg#adnabyddwr-endidau-ner
FlexiTerm Cymraeg	https://github.com/ispasic/FlexiTermCymraeg
FreeTxt	https://www.freetxt.app/
Gwyneth and Geraint	https://www.rnib.org.uk/nations/walescymru/welsh-synthetic-voices/
Helo Blod	https://businesswales.gov.wales/heloblod/
Helo Blod resource list	https://businesswales.gov.wales/heloblod/welsh-language-technology
Hir-iaith	https://www.hiriaith.cymru/hilite
Hunspell	https://github.com/techiaith/hunspell-cy https://github.com/techiaith/hunspell-cy-llafar
Hwb	https://hwb.gov.wales/
Hwb Resources	https://hwb.gov.wales/resources/?hideTabHeader=1
Lemmatizer	https://techiaith.cymru/cloud/api/lemmatizer/?lang=en
Lleisiwr	https://lleisiwr.techiaith.cymru/?lang=en
Macsen	https://macsen.techiaith.cymru/#/
MapioCymru	https://openstreetmap.cymru/
Meddal.com	https://www.meddal.com/meddal/
Micro:bit	https://microbit.org/code/
Named entities	https://hypermedia.research.southwales.ac.uk/kos/wnlt/, https://hypermedia.research.southwales.ac.uk/kos/wnlt2/
Normalization component	https://github.com/techiaith/techiaith-tts
Noun chunker	https://github.com/techiaith/talpiwr_enwol_cymraeg
Open Translation Memories	https://cofion.techiaith.cymru/en/search
OpenAI ChatGPT	https://openai.com/chatgpt
OpenAI GPT-4	https://openai.com/gpt-4
Paldaruo	https://techiaith.cymru/data/corpora/paldaruo/?lang=en
Part of speech tagger	https://github.com/techiaith/spacy-tagiwr-ency
Recording script for voice talents	https://techiaith.cymru/data/corpora/recording-script-for-voice-talents/?lang=en
Sentiment Analysis	https://github.com/cardiffnlp/en-cy-bilingual-embeddings
Si-Lwli	https://www.silwli.cymru/en/
SpaCy Welsh model	https://github.com/techiaith/spacy_cy_tag_lem_ner_lg
Speech model	https://huggingface.co/techiaith/wav2vec2-base-cy

Resource	Link
Stemmer	https://hypermedia.research.southwales.ac.uk/kos/wnlt/ https://github.com/CorCenCC/WelshStemmer
Techiaith	https://techiaith.cymru/?lang=en
Technocamps	https://www.technocamps.com/en/
Text-to-speech voices	https://tts.techiaith.cymru/#
Thesaurus	https://digigrid.cymru/reference/
Tiwtiadur	https://ytiwtiadur.corcencc.org/index.php?language=CY
Transcriber	https://trawsgrifiwr.techiaith.cymru/
Transcript Bank	https://git.techiaith.bangor.ac.uk/data-porth-technolegau-iaith/banc-trawsgrifiadau-bangor
Training sentence prompts corpus	https://techiaith.cymru/data/corpora/brawddegau-sentences/?lang=en
Trio Writing	https://digitalpublicservices.gov.wales/meet-user-needs/how-do-trio-writing
University of South Wales stopwords	https://hypermedia.research.southwales.ac.uk/kos/wnlt/
Vocab	https://cbsc04.techiaith.cymru/vocab-js-dev-demo/LlywodraethCymru/Addysg_a_sgiliau.html
Welsh punctuation and capitalisation model	https://github.com/techiaith/docker-atalnodi-server
Wicipedia	https://cy.wikipedia.org/wiki/Hafan
WordNet Cymraeg	https://github.com/CorCenCC/wncy