Roadmap for a Welsh-Breton Linguistic AI Network

Dewi Bryn Jones, Sasha Wanasky, Gweltaz Duval-Guennoc, Preben Vangberg, Leena Farhat, Mélanie Jouitteau, Loïc Grobol, Delyth Prys

March 2025

Published as part of the Agile Cymru Welsh Government funded project to establish a Welsh-Breton Linguistic AI network, project no. W10284

An Agile Cymru grant by the Welsh Government funded a three-month project in early 2025 to establish a Welsh – Breton Linguistic AI network to facilitate cooperation between researchers in the two countries. Welsh and Breton are closely related languages, sharing many linguistic peculiarities such as initial consonant mutations which affects how they need to be manipulated in digital contexts. Both are minoritized languages, aiming to make the best use of the latest AI technologies to serve their linguistic communities and achieve language revitalization and economic regeneration. The network wishes to share insights, best practice, resources and future collaborations to achieve common aims, and both current state of play and short term and longer-term aspirations are set out in the network's Roadmap. Speech recognition, text to speech and machine translation activities are described, including prototyping transcription models, and fine tuning/domain adaptation to improve support for named entities. Details of forthcoming developments are given, involving transcription tools, and large language models. Taking a strategic view, the importance of large amounts of appropriately licenced, annotated and tagged data is noted, together with the development of common standards and the need to ensure longer term sustainable repositories. Future needs include extending research and applications from the current emphasis on the standard language to other language varieties, including dialects and informal registers, especially to help the creative and media sectors with the provision of dedicated digital tools. Other economic sectors which could benefit from improved AI for Welsh and Breton include software and games, health and language learning, and translation. Bilingual provision can help expansion into multilingual markets. Longer-term investment and cross-region projects are highlighted for cost-effectiveness and the better pooling of resources. The network is committed to seeking joint funding opportunities for their common aims.

Map Ffordd Rwydwaith Deallusrwydd Artiffisial Ieithyddol Cymraeg-Llydaweg

Drwy gymorth grant Agile Cymru gan Lywodraeth Cymru ariannwyd project tri mis yn gynnar yn 2025 i sefydlu rhwydwaith Deallusrwydd Artiffisial Cymraeg – Llydaweg i hyrwyddo cydweithio rhwng ymchwilwyr yn y ddwy wlad. Mae Cymraeg a Llydaweg yn ddwy iaith sy'n perthyn yn agos i'w gilydd, yn rhannu nifer o nodweddion ieithyddol megis treigliadau cytseiniol ar ddechrau geiriau sy'n effeithio ar sut y dylid eu trin mewn cyd-destunau digidol. Mae'r ddwy yn ieithoedd lleiafrifedig, sy'n ceisio gwneud y gorau o'r technolegau DA diweddaraf i wasanaethu eu cymunedau ieithyddol ac adfywio eu hieithoedd a'u heconomi. Dymuna'r rhwydwaith rannu mewnwelediadau, arfer gorau, adnoddau a chynlluniau ar y cyd er mwyn gwireddu eu dyheadau, ac amlinellir y sefyllfa bresennol ynghyd â bwriadau tymor byr a thymor hirach ym Map Ffordd y rhwydwaith. Disgrifir gwaith adnabod lleferydd, testun i leferydd, a chyfieithu peirianyddol, gan gynnwys prototeipio modelau trawsgrifio, a mireinio/addasu parth i wella enwi endidau. Rhoddir manylion datblygiadau sydd i ddod, gan gynnwys offer trawsgrifio, a modelau iaith mawr. Gan gymryd golwg strategol, nodir pwysigrwydd meintiau mawr o ddata trwyddedig, eu hanodi a'u tagio'n briodol, ynghyd â datblygu safonau cyffredin a'r angen i sicrhau storfeydd cynaliadwy i'r tymor hirach. Ymhlith anghenion i'r dyfodol mae ehangu ymchwil a chymwysiadau o'r pwyslais cyfredol ar yr iaith safonol i amrywiadau eraill ar iaith, gan gynnwys tafodieithoedd a chyweiriau anffurfiol, yn arbennig i helpu'r sectorau creadigol a chyfryngau drwy ddarparu offer digidol pwrpasol. Ymhlith sectorau eraill o'r economi a allai elwa ar well DA ar gyfer y Gymraeg a'r Llydaweg mae meddalwedd a gemau, iechyd a dysgu

iaith, a chyfieithu. Gall darpariaeth ddwyieithog helpu ehangu i farchnadoedd amlieithog. Tanlinellir buddsoddiad tymor hirach a phrojectau traws-ranbarthol er mwyn bod yn gost-effeithiol a rhannu adnoddau yn well. Mae'r rhwydwaith wedi ymrwymo i chwilio am gyfleoedd ariannu ar y cyd ar gyfer eu hamcanion cyffredin.

Follenn-hent ar rouedad IA yezhoniel kembraek-brezhonek

Gant ur yalc'had Agile Cymru diwar gouarnamant Kembre e oa bet arc'hantaouet ur raktres tri miz e penn-kentañ 2025 evit sevel ur rouedad IA yezhoniel kembraek – brezhoneg evit aesaat ar c'henlabour etre enklaskerien an div vro. Yezhoù tost-tre eo ar c'hembraeg hag ar brezhoneg, meur a elfenn yezhel dibar dezho evel ar c'hemmadurioù a levezon an doare ma ranker ober ganto e kenarroudoù niverel. Yezhoù minorelaet eo an div, o klask tennañ o mad deus an teknologiezhioù IA diwezhañ evit servijout o c'humuniezh yezh, ha buhezekaat ar yezhoù kement hag an armerzh. Hon rouedad a lak da bal rannañ ar mennozhioù, an doareoù gwellañ d'ober, ar binviji hag ar c'henlabourioù da zont evit tizhout hon palioù boutin. Follenn-hent ar rouedad a zo enni ar stad a-vremañ hag ar c'hoantoù war berr ha war hir dermen. Deskrivañ a reomp an oberezhioù anaoudegezh ar gomz, sintezenn ar gomz hag an droidigezh emgefre, en o zouez ar sevel pimpatrom evit an treuzskrivañ, hag an affinañ / azasaat ouzh an domani evit gwelloc'h kemer e karg an elfennoù anvet. Deskrivañ a reomp dre a munud an diorroadurioù a rakwelomp evit a-sell ouzh an ostilhoù treuzskrivañ hag ar patromoù ledan yezh. Gant ur sell strategiel, e notennomp eo pouezus ar c'hementadoù bras a roadennoù gwirioù digor dezho, notennet ha tikedennet a-feson, asambles gant standardoù boutin, hag an ezhomm a vernadurioù padus war an hir dermen. E-touez an ezhommoù da zont, e notenner astenn an enklaskoù hag an arloadoù dezho da gemer e kont gwelloc'h ar rannyezhoù hag liveoù yezh disheñvel, en tu all deus hon rannyezhoù unvan, dreist-holl evit sikour rannadoù ar c'hrouiñ hag ar mediaoù da ginnig ostilhoù niverel gouestlet. Tachennoù ekonomikel all hag a c'hallfe tapout gounid eus an IA gwellaet evit ar c'hembraeg hag ar brezhoneg a zo ar meziantoù hag ar c'hoarioù, ar yec'hed, ar c'helenn yezh, hag an droidigezh. Pourvezadurioù divyezhek a c'hell sikour an emled war ar marc'hadoù liesyezhek. Lakaat a reomp war-raok ar raktresoù arc'hantaouiñ war hir dermen ha diouzh ar rannvroioù evit ar c'houst-efedusted hag ar gwellrannañ eus an traoù. Emouestlet eo hon rouedad da glask doareoù arc'hantaouiñ boutin evit hon kenbalioù.

Feuille de route du réseau d'IA linguistique gallo-breton

Une subvention Agile Cymru du gouvernement gallois a financé un projet de trois mois début 2025 pour établir un réseau d'IA linguistique gallois-breton afin de faciliter la coopération entre les chercheurs de nos deux pays. Le gallois et le breton sont des langues étroitement liées, partageant de nombreuses particularités linguistiques telles que les mutations des consonnes initiales, ce qui affecte la façon dont elles doivent être manipulées dans des contextes numériques. Toutes deux sont des langues minorisées, qui cherchent à utiliser au mieux les dernières technologies de l'IA pour servir leurs communautés linguistiques et servir la revitalisation de la langue et le soutien économique. Notre réseau souhaite partager ses connaissances, ses meilleures pratiques, ses ressources et ses collaborations futures afin d'atteindre nos objectifs communs. La feuille de route du réseau présente à la fois l'état des lieux actuel et les perspectives à court et à long terme. Nous

décrivons les activités de reconnaissance vocale, de synthèse vocale et de traduction automatique, y compris le prototypage de modèles de transcription et l'affinage / adaptation au domaine dans le but d'améliorer la prise en charge des entités nommées. Nous donnons des détails sur les développements à venir concernant les outils de transcription et les larges modèles de langues. D'un point de vue stratégique, nous notons l'importance de grandes quantités de données de droits d'utilisation ouverts, annotées et étiquetées, ainsi que le développement de standards communs et la nécessité d'assurer des dépôts durables sur le long terme. Nous identifions parmi les besoins futurs l'extension de la recherche et des applications au-delà du focus actuel sur la langue standard à d'autres variétés de langues, y compris dialectales et de registres informels, en particulier pour aider les secteurs de la création et des médias par l'offre d'outils numériques dédiés. D'autres secteurs économiques pourraient bénéficier de l'amélioration de l'IA pour le gallois et le breton, notamment les logiciels et les jeux, la santé et l'apprentissage des langues, ainsi que la traduction. La capacitation bilingue peut aider l'extension à des marchés multilingues. Nous mettons en avant les investissements sur le long terme et les projets interrégionaux pour leur rentabilité et une meilleure mise en commun des ressources. Notre réseau s'engage à rechercher des possibilités de financement conjoints pour nos objectifs communs.

1 Project Background....6
2 Linguistic Peculiarities of Welsh and Breton for Digital Manipulation....7
3 Overview of Current State of Play of Digital Resources for Welsh and Breton....9
4 Report of Pilot Project....12
4.1 Replicating Real-Time Breton Speech Model for Welsh....12
4.2 Lessons Learned from Cymen's Speech Transcription Project....13
4.3 Named-Entity Enriched Real-time Models....14
4.4 User Interfaces....15
4.5 Machine Translation....16
5 Workshop Reports and Visits....16
5.1 Report on Bangor Workshop....16
5.1.1 Visit to Cymen Translation Company....16
5.1.2 Visit to M-Sparc Science Park....17
5.1.3 Report on Roundtable Discussions....18
5.2 Report on Visit to Brittany....19
5.2.1 Visit to Breton School....19
5.2.2 Visit to Dizale....19
5.2.3 Roundtable Discussion....20
5.2.4 Media Coverage of Breton Visits....21
6 Future Direction of Work....21
6.1 Extending Current Activities....21
6.1.1 Speech Recognition....21
6.1.1.1 Improved Training Environment....21
6.1.1.2 Multi-Style Transcription....22
6.1.1.3 Pre-transcriptions Website for Linguists....22
6.1.1.4 Anaouder Graphical Transcription Software....22
6.1.1.5 Fine Tuning and Adaptation....23
6.1.1.6 ANR Project YAR....23
6.1.2 Text-to-Speech....24
6.1.3 Large Language Models....25
6.1.3.1 Breton and Welsh Chat Models....25
6.1.3.2 Speech Large Language Models....25
6.1.4 Applications....25
6.1.4.1 Speech Enabled Keyboards....26
6.1.4.2 Voice Assistant....26
6.2 Strategic and Long-term Considerations....26
6.2.1 Data....26
6.2.2 Empirical Scope....27
6.2.3 Economic Dimensions....27
6.2.4 Funding....28
7 Reflections and Future Work....29

1 Project Background

Wales and Brittany have a long history of mutual contact and cooperation. The Welsh and Breton languages are closely related, both belonging (alongside Cornish) to the Brythonic branch of the Celtic language group. There has been a long history of cultural interchange, cemented in more recent times with the signing of a Memorandum of Understanding between The Regional Council of Brittany and the Welsh Government in 2004. This was renewed in 2023[1], with areas of cooperation given as: Culture, Language policy, Tourism, Natural and cultural heritage, Sustainable development, Teaching and research, Youth, sport and European citizenship, Agriculture, agrifood and rural development, Economy and innovation, Cyber security, and Health, social policy and equality.

An initiative called Agile Cymru was launched by the Welsh Government, to encourage joint projects between Wales and Brittany[2]. It is this initiative which has funded the establishment of a Welsh-Breton Linguistic AI network, and we thank the Welsh Government for the funding and support we have received. Our project is unique in that it encompasses Language Policy, Teaching and research, and Economy and innovation in its activities.

In early July 2023 Welsh and Breton researchers in language technologies and linguistic AI made contact at the Gouel Broadel ar Brezhoneg in Langoned, Brittany, where Welsh researchers had been invited to present their work by Renan Kerbiquet, Chair of the board for Mignoned ar Brezhoneg, which was organizing the festival. Following this, Mélanie Jouitteau and Loïc Grobol were invited to present their research on Breton Language Technologies at Bangor University in their Welsh Language Technologies Academic Symposium in December 2023. Their contributions were consequently published in Language and Technology in Wales: Volume 2 [Gareth Watkins et al, Bangor University 2024][3].

Another event worthy of note was the workshop organised by IKER Laboratory (CNRS) at Quimper University in June 2024, to facilitate a meeting of minds between linguists and developers of technologies for Breton and Brythonic languages, foster a deeper understanding of each other's achievements, and to build a collective capacity in this field. Several Bangor students had their paper accepted at the conference[4].

These early attempts at Welsh-Breton collaboration led to a deeper desire to work together, and in December 2024 the Language Technologies Unit at Canolfan Bedwyr, Bangor University, Wales, received a small grant from the Agile Cymru fund to establish a Welsh-Breton Linguistic AI Network. The aim of the network is to focus on knowledge exchange between speech and language experts working on language centric artificial intelligence for both communities.

---

[1] https://www.gov.wales/cross-cutting-bi-lateral-agreements

[2] https://www.gov.wales/initiative-encourage-joint-projects-between-wales-and-brittany-agile-cymru-html

[3] https://research.bangor.ac.uk/portal/files/75338920/Language_and_Technology_in_Wales_-_Volume_II_-_ISBN_978-1_84220-207-4.pdf

[4] https://arbres.iker.cnrs.fr/index.php?title=2024_Workshop_on_Breton_Language_Technologies

Language revitalization through language technology has long been a theme of researchers in both countries, and as such, our activities may be viewed as a sub-area of Language Policy. As university-based researchers we naturally have a focus on teaching and research, and we take particular pride in being the recipient of a scholarship provided by Cymen (a translation company based in Caernarfon) to our MSc. Language Technology Program is Alan Kersaudy, a Breton who is fluent in both Breton and Welsh. Although we naturally enjoy the great cultural ties that Wales and Brittany enjoy, the focus of our research is the field of Conversational AI, especially where it can help communication between humans and computers through the medium of the closely related Welsh and Breton languages. Research and development of language centric AI in general, and especially conversational AI for Welsh and Breton, have a great potential for contributing to the economy of both Wales and Brittany, directly through technological advancements by SMEs based in our areas, and indirectly through cutting costs and enabling fuller participation to various sectors.

Hubs such as the M-Sparc Science Park on Anglesey have impacted positively on the software, technology and creative sectors in Wales, whilst at the same time strengthening the Welsh language and culture. Such successes deserve to be more widely known and emulated, and it is for this reason that we have established this new Welsh-Breton network. It is hoped that the network will continue long after the initial funding period, and that it will be a springboard to develop other joint projects together.

In addition to Bangor University, project partners include researchers at the Université de Pau et des Pays de l'Adour & Université Bordeaux-Montaigne (CNRS IKER Lab), the University of Bretagne Ouest (UBO), and Université de Nanterre, Paris, and the Gouel Broadel ar Brezhoneg (Breton Language Festival) which facilitated the first meeting between Linguistic AI researchers in Wales and Brittany, and the Cymen translation company in Caernarfon, Wales, and M-SParc Menai Science Park on Anglesey, Wales.

2 Linguistic Peculiarities of Welsh and Breton for Digital Manipulation

Much of the initial development of digital computers happened in Britain and the US, and therefore English became the dominant language in the digital sphere. Consequently, conventions associated with English and the Latin script became the default in the digital sphere, even to the extent that, for example, only domain names conforming to English-based ASCII character set were accepted as recently as 2003. That nearly all programming languages are English-based is a separate issue and is not under discussion in this Roadmap. The issue discussed here concerns the ease of development of language-based digital applications, the creation and curation of digital content, and the technical support for them. It also concerns linguistic AI for non-English languages, including the creation of Large Language Models (LLMs) and other resources, especially for less-resourced and minoritized languages such as Breton and Welsh.

Welsh and Breton, being Brittonic Celtic languages, share several linguistic features that pose challenges for language technologies. Both languages feature complex systems of consonant mutations, where the initial consonant of a word changes depending on its grammatical context. This

creates significant variability in word forms, making it difficult for algorithms to recognize and process words consistently.

Unlike many widely used languages that follow Subject-Verb-Object (SVO) order, Welsh and Breton typically use VSO order. This difference in sentence structure requires language technologies to be adapted to handle this less common syntax. This difference in syntax creates complications for parsing and machine translation.

Both languages also use inflected prepositions, where prepositions combine with pronouns to form single words. This adds another layer of complexity to morphological analysis. Welsh and Breton also use verbal nouns, which function as noun forms of verbs instead of infinitives. This structural difference requires specialized processing.

As described in the Language Report on the Welsh language, contained in the book *European Language Equality: A Strategic Agenda for Digital Language Equality*[5], the Welsh alphabet contains 29 letters, including eight digraphs (e. g., ch) and the letter j borrowed from English to represent the borrowed /dʒ/ consonant phoneme. V, x and z are not used in Welsh, but are included with the alphabet for computer use as they often appear in named entities such as foreign place names..... Accent characters are common over vowels. Welsh has a continuum of other registers, with colloquial or informal registers differing markedly from the standard written form. It has many local dialects, with the main difference between those of north and south Wales. Welsh has two methods of verb formation, utilising concise forms or periphrastic forms, using auxiliary verbs. Guidelines to the latest version of the modern Welsh orthography, first standardised in 1928, were published in 1987 by the Literature Panel of the Board of Celtic Studies. In 2021 a new Welsh orthography Panel was established by the Welsh Government, which aims to resolve minor inconsistencies in the orthography.

There was no individual Language Report on Breton in the *Language Equality* publication, and a comparable description of Breton for digital manipulation is yet to be written. However, briefly, Breton has 25 letters in its alphabet, according to the *Peurunvan* [unified] orthography, namely a, b, ch, c'h, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, t, u, v, w, y, z, and the following accented characters: â, ê, î, ô, û, ù, ü, ñ. Like Welsh it has consonant mutations at the beginning of words under certain grammatical conditions. It also has many different language registers from the very formal to very informal, and markedly different dialects in Lower Brittany where it has traditionally been spoken: Kerne, Leon, Goelo, Treger in the west of Brittany and Gwenedeg in the south-east. Standard Breton also emerged in the 20th century, and some spoken Neo-Breton varieties in young adults and children has some distinct prosodic properties [[6], [7]]. There are some Breton competing orthographies, with

[5] *European Language Equality: A Strategic Agenda for Digital Language Equality* (eds. Rehm and Way 2023) Springer, pp 223-226
https://link.springer.com/content/pdf/10.1007/978-3-031-28819-7.pdf?pdf=button

[6] More Information on Breton varieties and their syntactic differences can be found in Mélanie Jouitteau. Standard Breton, traditional dialects, and how they differ syntactically. Journal of Celtic Linguistics, 2020, 21 (1), pp.29-74. https://hal.science/hal-02269458v1

[7] About the continuity of prosodic structure in different generations of natives, see Elfner, Emily, Francesc Torres-Tamarit & Mélanie Jouitteau. 2024. 'Prosodic phrasing in Breton', (éds.), Minghui Dong, Yanfeng Lu, and Ying Chen (éds.), *East Meets West: Explorations in Tone and Intonation*, Proceedings of the Second International Conference on Tone and Intonation, November 18–20, 2023, Chinese and Oriental Languages Information Processing Society. *Proceedings of TAI 2023*, 77-78.

the *Peurunvan* orthography being centrally recognised and taught in Breton-medium schools. A standard also has emerged for the Vannetais dialect. Resources can still be found in other orthographies like Skolveurieg, but they are used neither in edition nor schools.

3 Overview of Current State of Play of Digital Resources for Welsh and Breton

Compared to widely spoken languages, Welsh and Breton have fewer digital resources, such as large text and speech corpora, which are essential for training language models. This lack of large datasets makes it more difficult to create accurate and robust language technology tools. This lack of data is exacerbated through the regional dialects of both languages, which add even more variation that needs to be accounted for in the data sets.

These features, combined with the relatively smaller number of speakers and available resources contribute to the challenges of developing effective language technologies for Welsh and Breton.

Nonetheless, language technology resources exist for both Welsh and Breton including resources such as tagged text corpora for a variety of purposes, POS-taggers and lemmatizers, and AI speech recognition and text to speech models. The European Language Grid websites offer a service to search the large majority of available resources.

Welsh: https://live.european-language-grid.eu/catalogue/?&language__term=Welsh

Breton: https://live.european-language-grid.eu/catalogue/?&language__term=Breton

Resources such as AI models and datasets to train AI model by organisations and individuals can also be found on Hugging Face by searching for language tags:

Welsh: https://huggingface.co/models?language=cy&sort=trending

Breton: https://huggingface.co/models?language=br&sort=trending

Other websites include the Entrelangues page for Breton which lists all Breton resources under the “Digital Resources” tab: https://entrelangues.modyco.fr/index.php/Breton

Most organisations and some individuals will also upload their AI models, datasets and code to GitHub.

Welsh: https://github.com/topics/welsh

Breton: https://github.com/topics/breton

In terms of datasets, there are currently around 204 hours of Welsh speech data that are being used to train Welsh speech recognition models and 89 hours of Breton data. The Welsh datasets are the large CommonVoice corpus of read speech which makes up 61% of all the data, the Banc Trawsgrifiadau and Lleisiau Arfor corpora of spontaneous speech which add up to 37% and the small Enwau Cymraeg data set which contains read speech with a large number of personal names and makes up 2% of all the available data. The Breton datasets consist of a larger variety of scraped recordings from the web and are not entirely open-source, unlike the Welsh datasets. The largest

amount of data comes from the BrezhoWeb Youtube channel and makes up 45% of the Breton data. The Breton Common Voice corpus amounts to around 30% of the data and the remaining 25% consisting of smaller Text to Speech and other linguistic corpora, as well as, more scraped recordings and Breton dubbed films. While Welsh contains more data overall, the proportion of read and spontaneous speech is more balanced in the Breton data. The Breton data, however, is to a large extent not open-source and cannot be shared online and for commercial purposes, while the Welsh data is entirely available under a CC-0 licence.

In June 2023, the European Language Equality (ELE) Initiative, together with the European Language Grid, as well as many other relevant stakeholders from the European Language Technology, Natural Language Processing, Computational Linguistics and language-centric AI communities, published their results on quantifying the level of technological support for all languages in Europe. Their findings proved that a large imbalance exists between English and a vast majority of other languages. Thus ELE's vision and subsequent goal strategic agenda is to accomplish digital language equality in Europe by 2030.

The following charts are extracted from the European Language Grid's Catalogue Dashboard (https://live.european-language-grid.eu/catalogue/dashboard) that highlight their findings on levels of support for Welsh and Breton. They highlight that both languages suffer from significant imbalances when compared to Irish, and respective dominant languages in our respective bilingual communities.

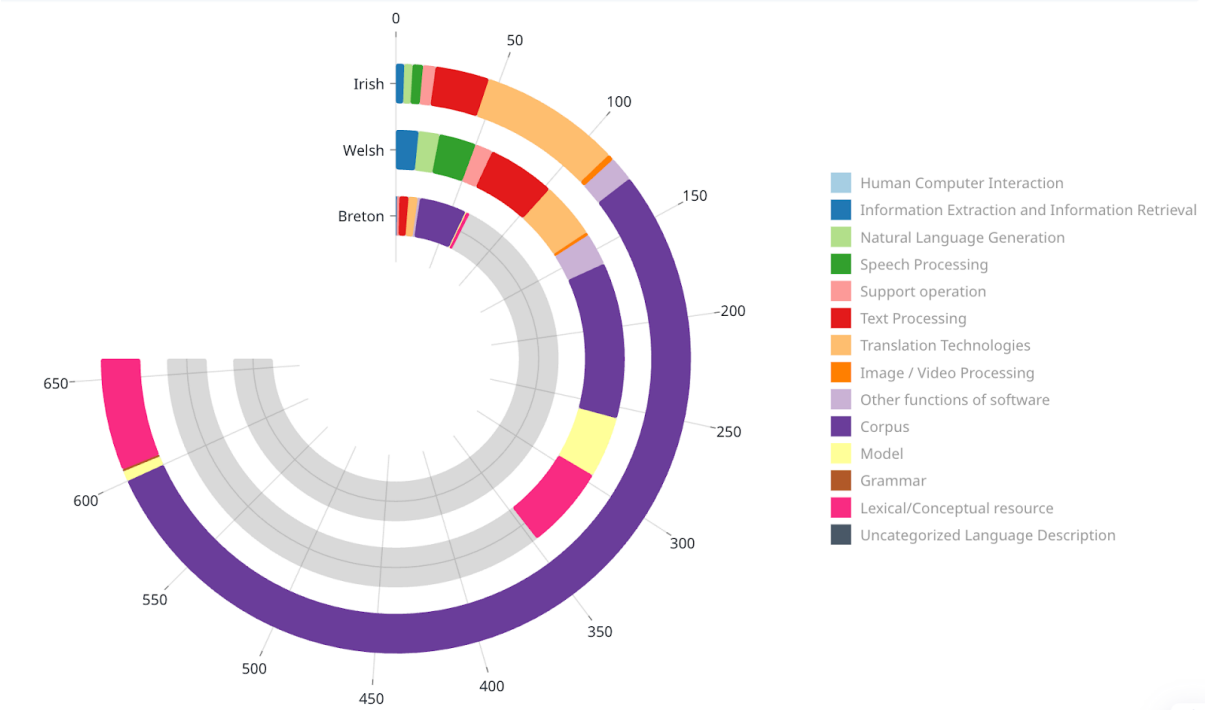


Figure 1 - comparison between language technology support for Irish, Welsh and Breton

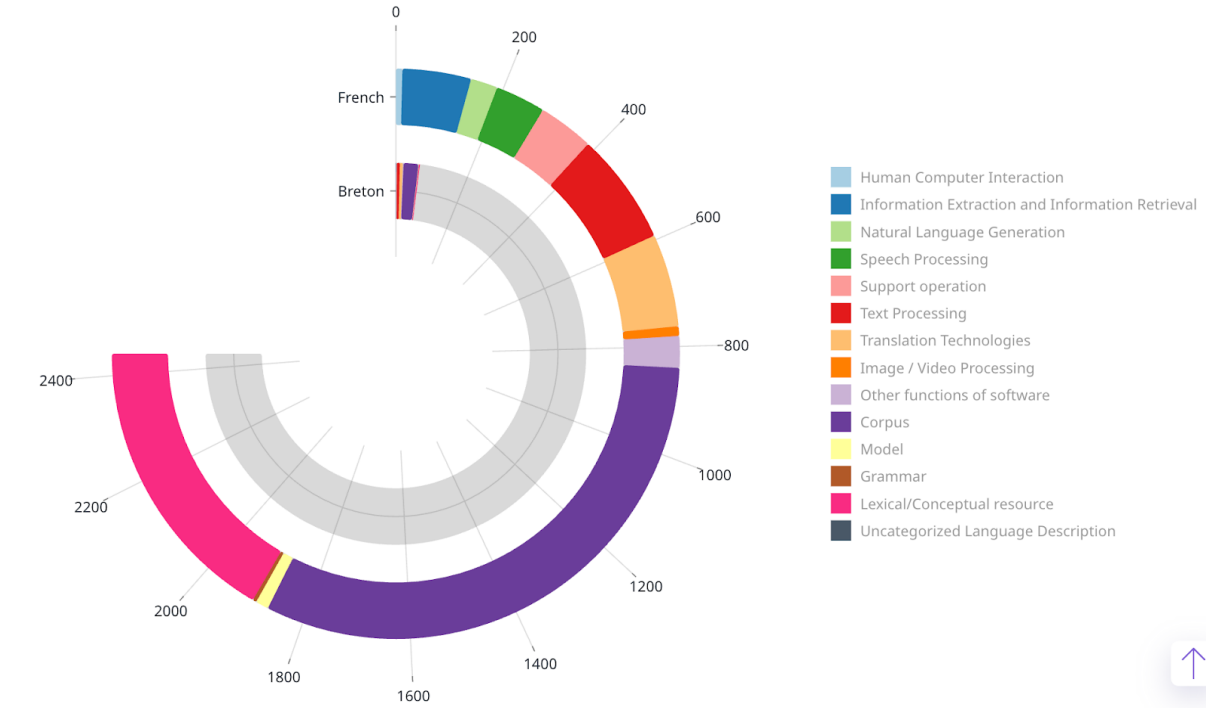


Figure 2 - comparison between language technology support for French and Breton

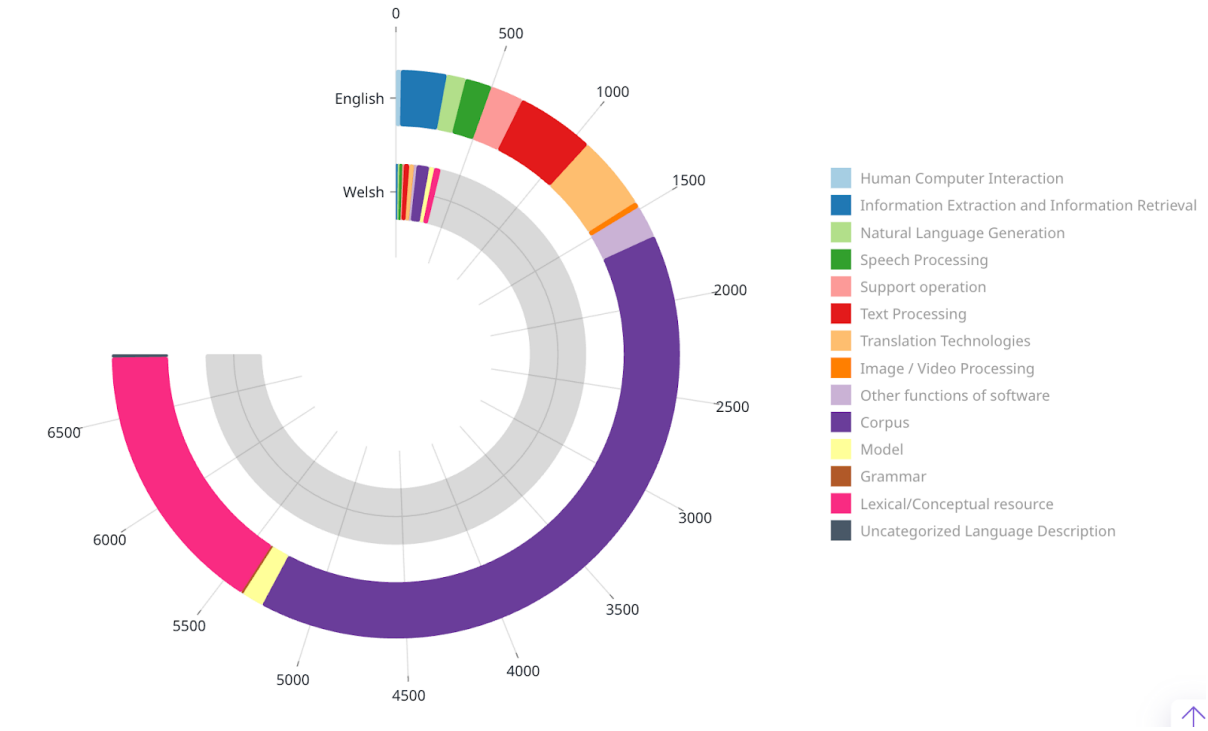


Figure 3 - comparison between language technology support for Welsh and English.

4 Report of Pilot Project

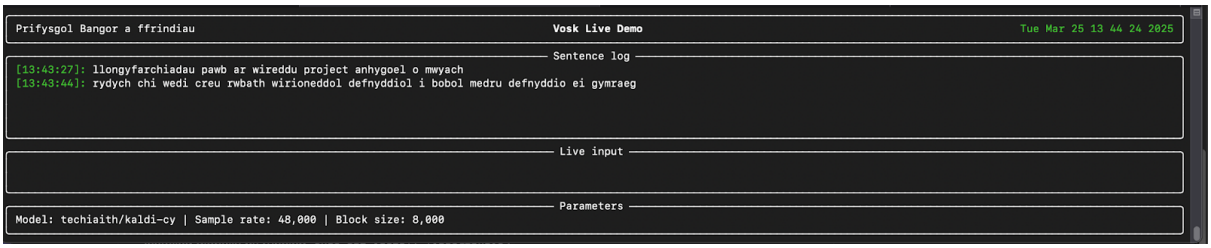
4.1 Replicating Real-Time Breton Speech Model for Welsh

Time was spent to replicate work undertaken to create real-time Breton speech models for Welsh. The result of this work was a series of initial models trained using a training environment using Docker. This environment hence is very portable and reproducible and and is freely available from GitHub[8] according to a permissive MIT open source license. Models that have been trained are freely available from the HuggingFace website[9].

A prototype application was also developed which can run on Linux and Mac OS machines. (N.B. we have not tested within Windows using Microsoft's Linux on Windows WSL). To install and get started transcribing on a suitable machine, open a Terminal Window and type:

```
$ git clone https://github.com/Cymru-Breizh-Agile-Cymru-Project/vosk-tui.git
$ cd vosk-tui
$ uv run vosk-tui -m techiaith/kaldi-cy
```

This will start the transcription application as well as download (if needed) the Welsh model from the Hugging Face website. As soon as you begin speaking (in Welsh), a fairly highly accurate text transcription appears immediately at the bottom of the application. Transcriptions from each utterance are added to the history in the central main window.



The hope is that this code can be used as a foundation to train further models for Welsh, Breton, as well as other languages.

The trained Welsh models themselves have a word error rate (WER) of about 34% on a test set consisting of verbatim transcriptions of spontaneous speech. The models perform better on read speech, which is to be expected. This is statistically significantly worse (*p<0.001*) than OpenAI whisper-large-v3 (~23%), and techiaith/whisper-large-v3-ft-commonvoice-cy[10] (~18), but statistically significantly better (*p<0.001*) than whisper-base (~57%) which is more comparable in size to the Kaldi-Vosk models. During the phase of models evaluation, network member Preben Vangberg created a new software library[11] for calculating the WER (and CER) error metrics that is magnitudes

[8] https://github.com/Cymru-Breizh-Agile-Cymru-Project/vosk-cymraeg
[9] https://huggingface.co/techiaith/kaldi-cy
[10] https://huggingface.co/techiaith/whisper-large-v3-ft-commonvoice-cy
[11] https://gitlab.com/prebens-phd-adventures/universal-edit-distance

faster than the library provided by the Hugging Face company and all other alternatives. This facilitated being able to produce a long and thorough list of evaluation results and comparisons as seen below.

	model	dataset	wmer	cmer	wer	σwer	cer	σcer
2	DewiBrynJones/evals-whisper-large-v3-ft-btb-ca-cy		23,87%	9,31%	25,39%	0,68	11,19%	0,53
3	DewiBrynJones/evals-btb-whisper-large-v3-ft-btb-cv-ca-cy	btb	22,73%	8,52%	23,48%	0,26	9,68%	0,21
4	DewiBrynJones/evals-btb-whisper-large-v3-ft-btb-cv-ca-cy-2502	btb	23,33%	8,88%	24,08%	0,40	10,04%	0,28
5	DewiBrynJones/evals-btb-whisper-large-v3-ft-btb-ca-cy	btb	23,57%	9,06%	24,40%	0,45	10,37%	0,37
6	prvInSpace/eval-btb-kaldi-full-model	btb	34,41%	18,15%	35,54%	0,28	19,58%	0,22
7	prvInSpace/eval-btb-kaldi-ner-text-only	btb	37,64%	20,31%	38,75%	0,28	21,70%	0,20
8	prvInSpace/eval-btb-kaldi-ner-full	btb	37,75%	20,57%	38,89%	0,29	22,18%	0,24
9	prvInSpace/eval-btb-kaldi-all-with-corpus	btb	37,95%	21,36%	38,38%	0,29	22,65%	0,24
10	prvInSpace/eval-btb-kaldi-bilingual	btb	41,88%	25,94%	43,11%	0,30	28,00%	0,26
11	DewiBrynJones/evals-btb-whisper-base-ft-btb-cv-cy	btb	57,19%	25,85%	55,78%	1,38	25,42%	0,83
12	DewiBrynJones/evals-ca25-whisper-large-v3-ft-btb-ca-cy	ca25	25,54%	11,33%	30,77%	0,52	14,96%	0,39
13	DewiBrynJones/evals-ca25-whisper-large-v3-ft-btb-cv-ca-cy	ca25	26,72%	12,41%	31,25%	0,76	15,45%	0,54
14	prvInSpace/eval-cvcy-kaldi-all-with-corpus	cv	19,42%	10,20%	21,43%	0,26	11,49%	0,18
15	prvInSpace/eval-cvcy-kaldi-bilingual	cv	24,65%	15,85%	26,77%	0,30	17,75%	0,26
16	prvInSpace/eval-cvcy-kaldi-full-model	cv	30,81%	12,21%	31,81%	0,26	12,61%	0,13
17	prvInSpace/eval-cvcy-kaldi-ner-full	cv	31,90%	13,33%	32,88%	0,26	13,81%	0,14
18	prvInSpace/eval-cvcy-kaldi-ner-text-only	cv	32,50%	13,63%	33,53%	0,27	14,12%	0,14
19	DewiBrynJones/evals-cv-whisper-base-ft-cv-cy	cv	38,68%	13,25%	39,85%	0,30	13,89%	0,37
20	DewiBrynJones/evals-cv-whisper-base-ft-cv-cy-en	cv	44,16%	15,04%	45,86%	0,51	15,93%	0,44
21	DewiBrynJones/evals-cvb-whisper-base-ft-cv-cy-en	cvb	35,73%	15,47%	37,75%	0,67	16,78%	0,71
22	DewiBrynJones/evals-ec-whisper-large-v3-ft-btb-cv-cy	enw	17,46%	4,96%	17,30%	0,17	4,99%	0,06
23	DewiBrynJones/evals-ec-whisper-btb-cv-ft-enwaucymru-lm	enw	17,77%	6,37%	17,56%	0,18	6,21%	0,08
24	prvInSpace/eval-enw-kaldi-ner-full	enw	32,04%	13,34%	32,36%	0,25	13,58%	0,13
25	prvInSpace/eval-enw-kaldi-all-with-corpus	enw	34,45%	21,40%	35,37%	0,27	22,79%	0,25
26	prvInSpace/eval-enw-kaldi-ner-text-only	enw	34,76%	15,26%	35,23%	0,25	15,47%	0,14
27	prvInSpace/eval-enw-kaldi-full-model	enw	38,46%	15,89%	38,58%	0,24	16,40%	0,14
28	prvInSpace/eval-enw-kaldi-bilingual	enw	47,05%	35,73%	48,82%	0,30	38,33%	0,32
29	prvInSpace/eval-lla-kaldi-full-model	lla	46,08%	25,26%	51,37%	0,34	30,71%	0,34
30	prvInSpace/eval-lla-kaldi-ner-full	lla	50,09%	28,62%	55,67%	0,33	34,96%	0,36
31	prvInSpace/eval-lla-kaldi-ner-text-only	lla	50,30%	28,80%	56,08%	0,33	34,96%	0,34
32	prvInSpace/eval-lla-kaldi-all-with-corpus	lla	82,49%	77,17%	80,03%	0,24	73,41%	0,26
33	prvInSpace/eval-lla-kaldi-bilingual	lla	88,74%	86,14%	87,52%	0,21	84,73%	0,23

None of the fine tuned Whisper models support transcribing in real-time which means that we have sacrificed accuracy in order to substantially reduce the latency thus making the models real-time.

But the hope is that there is more room for improvements in accuracy. For example, the models were not trained with extended language models, the hyper-parameters were not optimised, the phone-set was not evaluated, etc. These are all areas that require further research and which could improve the accuracy of the models.

4.2 Lessons Learned from Cymen's Speech Transcription Project

Beginning in January 2024, the translation company Cymen launched a pilot project to collect as much speech data as possible within a year. The goals of this project were to develop strategies and methods to improve the efficiency and quality of spontaneous speech data collection. The project was in collaboration with Bangor University's Language Technologies Unit and funded by Project Arfor[12]. Project Arfor sees the relationship between the Welsh language and the local community as mutually beneficial and through their challenge funds seeks to support businesses in the Arfor area in order to keep especially young and Welsh speaking people in the area. The large challenge fund was awarded to Cymen to address the lack of Welsh speech data, which is needed to improve the

[12] https://www.rhaglenarfor.cymru/

speech recognition models. The grant was used to employ a full time project officer to carry out the task of data collection and to employ a number of freelance transcribers to support them in the mammoth task of transcribing the collected recordings. After a year, Cymen successfully created the 35 hour transcribed audio dataset Lleisiau Arfor and a Windows App Trawsgrifiwr Cymen, which uses a model trained on the Lleisiau Arfor data to transcribe audio files.

Apart from the dataset the output also includes reports on lessons learned during the project and can be summarised as follows:

1. Most of the budget should go towards high quality and consistent transcription, since they need to be done by trained professionals and are very labour intensive.
2. Clear and detailed transcription guidelines should be developed to ensure consistency across different transcribers. These should also be adapted during the course of the project as and when issues arise.
3. Email templates to ask content creators or contributors for their consent to use their existing content should be created as well as consent forms for any newly created data.
4. Collaboration with event organisers and organisations that hold regular talks are very beneficial as these organisers can take on some of the administrative work such as asking for consent and recording the talks.
5. When recording conversations between volunteers, they should be discouraged to discuss any personal information. Conversation topics that discourage this further can be suggested if necessary.
6. More research should be done on identifying data gaps and biases in the existing data and models. One way of doing this is by collecting metadata on the speakers' accent, proficiency, socio economic status etc. This kind of data, however, also comes with its own issues since it needs to be kept safe and participants need to stay anonymous.

4.3 Named-Entity Enriched Real-time Models

One of the challenges in ASR models is that named-entities (such as people's names and place names) are often taken out of datasets due to privacy reasons. That causes the model performance to decrease for named entities.

To try to overcome this, we used a named-entity enriched dataset collected by network member Sascha Wanasky called 'Enwau Cymru'[13] to train a named-entity enriched real-time model. The preliminary results showed that adding this data statistically significantly improved the results for named-entities. Interestingly enough, smaller improvements could be observed when training with only the text from the 'Enwau Cymru' dataset. In theory therefore, the language model, which is responsible for generating text, can still learn without audio, how to correctly spell names.

Since language models in Kaldi can be trained separately and independently of acoustic modelling (in contrast to Whisper where some form audio must always be associated with the text for improving

[13] https://huggingface.co/datasets/wanasash/enwaucymraeg

its language modelling) we experimented with increasing the amount of texts enriched with named entities.

Bangor University's Language Technologies Unit has an internal archive[14] of BBC Radio Cymru news programs that have been automatically transcribed using the best fine tuned Whisper model. The Unit also has developed a named-entity recognizer[15] model in previous projects which could be put to use to extract the sentences containing people names that could be added to text corpora training. Unfortunately, the model discovered too few examples of correctly spelled names in the automatic transcriptions. It did however find a lot of mis-transcribed named entities which could be the basis for future research that would aim to conclude for sure whether larger amounts of text only data can improve the transcription of named entities (or could even make specialization or fine tuning of Kaldi-Vosk based speech recognition easier than Whisper)

Network member Gweltaz Duval-Guennoc in the meantime started the first steps to create a larger text only training resource for Breton named entities by anonymising and then enriching a dataset by swapping named-entities randomly. With this goal, a dataset was collected from a wikipedia dump[16], and pre annotated using a rule-based tagger. The dataset was then used to train a prototype Breton spaCy model[17], able to tag named entities automatically in any Breton sentence.

4.4 User Interfaces

Two prototype user-interfaces were developed during the project. Already mentioned in section 4.1, a simple terminal user interface (TUI) was developed for the purposes of demonstrating the real-time transcription capabilities of the models. This application works with any real time model based on the Kaldi-Vosk software and is available on the project's GitHub page.

The second was an online web-frontend for Gweltaz Duval Guennoc's Anaouder software and the Breton models. This is a very simple website which allows the models to be available to a larger audience, especially by less technical people. Any size media file, such as a recording or a video can be uploaded to the website and after a couple of minutes, automatic transcriptions are available for download in any standard format for editing and correcting in any suitable tool.

The website is also easily extendable, and the back-end can easily be replaced to support different models and languages. The code for both server[18] and web browser[19] components are available from the project's GitHub. A version is currently hosted by Bangor University and available at https://stt-br.techiaith.cymru/.

[14] Bangor University is not able to share and redistribute this data. The data also is not used to train any models that are subsequently published for third party developers and users.
[15] https://github.com/techiaith/spacy_cy_tag_lem_ner_lg
[16] https://huggingface.co/datasets/gweltou/wikipedia-br-20240325
[17] https://github.com/Cymru-Breizh-Agile-Cymru-Project/br-ner
[18] https://github.com/Cymru-Breizh-Agile-Cymru-Project/anaouder-server
[19] https://github.com/Cymru-Breizh-Agile-Cymru-Project/anaouder-webclient

4.5 Machine Translation

If the output of the new Kaldi-Vosk ASR model is provided as input to a machine translation model that is capable of translating from Welsh into English, then we have the beginnings of a speech to translated text capability.

Current support in OpenAI's whisper model achieves only a BLEU score of 1.53% when translating spontaneous Welsh language speech into English transcriptions[20].

Based on the work done for training a Breton to French model by fine tuning open source multilingual machine translation models pre-trained by Meta AI, Bangor also fine tuned a NLLB (No Language Left Behind) foundation model[21]. Bangor University had already ensured in other projects that sufficient Welsh-English translation data, collected from open licensing sources in the Welsh public sector, exists for experimentation and prototyping.

We created a prototype pipeline using the kaldi-full model and followed by the translation model. The result was a BLEU score of 25% which, while still having room for improvement, is significantly better than the OpenAI whisper model.

5 Workshop Reports and Visits

5.1 Report on Bangor Workshop

5.1.1 Visit to Cymen Translation Company

The network members were greeted warmly at the company's office in Caernarfon by the two directors and given a short history of the company and its use and development of language technologies. Their use of translation memory was of particular interest due to its size and incorporation into the day to day workflow of their translators. They started collecting and building the memory in 1999, which has exploded in size in recent years with around 400,000 words now being added every month. Due to its size, they were able to create isolated memories for long-term clients that require a particular translation style. Their investment into technology allowed them to stay ahead of the game by developing their own machine translation model instead of depending on global players such as Google. This data sovereignty also allowed them to remain in control of their own data instead of feeding it into Google's translation memory and jeopardise their clients' sensitive information.

[20] https://huggingface.co/datasets/techiaith/banc-trawsgrifiadau-bangor/viewer/default/translations
[21] https://huggingface.co/techiaith/nllb-200-1.3B-ft-cym-to-eng

With the help of funding by Arfor, Cymen ventured into speech recognition technology which would help them provide subtitling work for clients. After a year of collecting data, the Language Technologies Unit at Bangor University were able to train a new speech recognition model using this data and Cymen developed a Windows application that will from now on be used in their administrators' day to day workflow. It is also available for free on Cymen's website.

While Cymen is keen to invest in language technologies, they also stressed the importance of keeping humans in the translation process. There is a consensus among Welsh translators not to use large language models (LLM) for translating text due to reasons including legal issues, workers' rights and the endangerment of the Standard of Welsh. Rather, they would use LLMs in quality assurance or to help people gain more confidence in using their Welsh, such as drafting emails.

During these conversations, the Breton researchers shared their own experiences with the translation sector in Brittany and their use of translation and speech recognition technology. According to their reports, the translation sector is in need of more organisation, for example, through the establishment of a translation association and the sharing of tools, such as machine translation models, with the individual translators. At this time, the demand for translation into Breton is small but more cooperation among freelance translators and with researchers developing translation technologies could improve the situation.

The company Dizale was mentioned to have started to invest in speech recognition technology to create subtitles for films and the newly approved project Yezh ar Vro will provide a larger speech corpora to further improve this technology.

5.1.2 Visit to M-Sparc Science Park

The network members visited the M-Sparc Science Park as part of their two day workshop, and were first given a presentation on the establishment and achievements of the science park and its unique position in Wales' economy. M-Sparc is an innovative Science Park located on the island of Anglesey in north-west Wales and is the first Science Park in Wales. It currently has 79 tenants, split between the Low Carb, Energy and Environment, ICT and Commercial sectors. It is however much more than a location for companies' offices, with a mission to ignite ambition and innovation to power a sustainable Wales. It sees itself as a 'three legged stool' aiming to be financially, environmentally and culturally sustainable. Supporting the Welsh language is an integral part of its remit, creating a community and acting as a catalyst, creating high value innovative careers in a rural area, connecting skills and innovation, industry and academia. It includes Welsh as an everyday language, and has fun activities for staff and tenants learning Welsh, such as yoga, ukulele lessons and BBQs to normalise the language. Many of the companies located at M-Sparc use Welsh as part of their business offerings, using it both as a marketing tool and for creating products and services. Welsh has also been used in bilingual and multilingual products and services, helping companies to expand their exporting and international potential.

A tour through the M-Sparc premises included visits to two tenant companies: Brandified and Animated Technologies. The company Brandified was founded by Neil Thomas who jointly won the Hac Iaith in 2024 with his product Cardiau Cosmos that is meant to help children use more Welsh at home. The Cardiau Cosmos are collectable cards that contain NFC technology, which allow children

to access a variety of audiobooks through a mobile app. When the product is released, they are hoping to enable children that do not speak Welsh with their parents to use Welsh at home by listening to these audio books and to interact with other children in Welsh when collecting or trading the cards. The principle follows that of other popular children's toys such as the tonies and can be adapted for Breton.

Animated Technologies is a creative visual digital company that has created a virtual reality language learning app to help learners develop conversational skills in an interactive digital environment which they demoed for us. Designed originally to help schoolchildren learn Welsh, it features an imaginary town of Aberwla with a supermarket and other amenities where students could practice their Welsh. The app can be localized to other languages and is an example of how Welsh language apps can be adapted for other markets. Both companies' could expand their product to the Breton market if they wanted to.

5.1.3 Report on Roundtable Discussions

The roundtable, chaired by Leena Sarah Farhat and minuted by Sasha Wanasky that was held in M-Sparc on the second day of the workshop allowed members of the public, local business owners and researchers in Language Technologies to ask questions to the panelists, who specialise in Welsh and Breton speech recognition.

The discussions brought several needs to the surface that revolve around closer cooperation between the Celtic nations and between disciplines, the need for more data as well as more research on data gaps and biases in the models, and ways of bringing these new technologies to the end users.

The topics that were raised around increased cooperation involved mostly the benefit of in-person events such as the first workshop in Bangor. The researchers from both Wales and Brittany were able to exchange tools and methodologies more efficiently this way while also learning about the wider context in which the development of speech technologies is situated and how the culture and political situation of the partner countries has influenced these developments. The in-person workshops were also beneficial for the longevity and sustainability of the research network through conversations that went beyond the sharing of tools but also allowed the researchers to build long lasting personal relationships. As an additional benefit, the network also enables the researcher from Wales to work closer with the European Union since Brittany is still part of the EU and on a larger scale, enables them to exchange knowledge with other researchers when there is a lack of such research in their own countries. Finally, increased cooperation should also be prioritised between the disciplines of Computer Science and Linguistics, since linguistic expertise is needed in identifying data gaps, transcription and adding rules to rule-based and statistical models such as Kaldi-Vosk.

The need for more data can be narrowed down thanks to the achievement of previous data collection projects that highlighted particular areas that need more research and resources. These areas are the data gaps and biases that are currently present in the speech corpora. To identify these gaps, the researchers would have to create balanced evaluation sets and train models with a variety of data to uncover over- and underrepresentation of certain accents, speaker groups and registers. The need for variation in data is a well known fact but more work needs to go into creating

scientifically sound evaluation procedures and high quality data for the evaluation sets. After the identification of data gaps more data to fill these gaps would need to be collected including more metadata to simplify the identification of gaps in the future. Finally, transcribers need to be trained to take on the work of annotating this data and clear transcription guidelines need to be developed to ensure consistency between transcribers. This work is highly labour intensive and remaining consistent in style is challenging when dealing with the natural variation in people's speech. That's why training courses or internships for transcribers would be beneficial in ensuring adequate training for this crucial work.

Finally, there is a need for all the work that is done by the researchers to be accessible to others by offering the code and data as open-source and developing technologies with the end users in mind. Everything that has been published on Welsh and Breton Speech Recognition so far has been open-source, but more importance should be placed on the upkeep of repositories and documentations and providing smaller AI models that can run locally. The availability of these will benefit the longevity of the network and sovereignty over the two partners' speech technologies so they do not become dependent on larger corporations outside of Wales and Brittany. AI sovereignty also benefits the local economy by creating partnerships with small businesses and creating jobs. These partnerships are important since the businesses can assist in utilising the research in their end products and deliver these to everyday people that otherwise would not be able to access the newly developed technologies. More efforts to develop end products are needed to ensure that the work done by the researchers is accessible to everyone and that these products are adequately advertised. Especially in a low resource language setting such as Wales and Brittany, the application of native language technologies can help create more opportunities for the speakers of these languages to use them.

All in all, the discussions highlighted a need for more cooperation between the Celtic nations and disciplines and more research into data gaps, a focus on accessibility and availability of existing technologies and the importance of sovereignty over the development of these technologies.

5.2 Report on Visit to Brittany

5.2.1 Visit to Breton School

The Lise Diwan Karaez is one of the two high-schools, and the largest, from the Diwan network of schools by immersion in Breton. It offers a weekly coding club through which students are supported in developing their programming and electronics skills, under the tutelage of Gweltaz Duval-Guennoc. The research partners were shown around the premises and introduced to some of the students and their project. One student, Yohann Tézé, has created a simple Android voice assistant and would like to extend the program to include Breton ASR and speech synthesis.

5.2.2 Visit to Dizale

During the last day of the Welsh team's stay in Brittany, a visit was made to Dizale, a Breton dubbing organisation. Established in 1998 the company primarily dubs films and series for children but also

other successful films and series such as the Terminator and Star Wars into Breton in their highly professional dubbing studio in Kemper. The company is well-connected with other minority languages and dialects in France and abroad and has begun to dub into Occitan and Gallo. Their future plans include closer cooperation with other Celtic languages to collectively obtain rights to large film and TV productions.

5.2.3 Roundtable Discussion

The first day of the workshop can be summarised as revolving around discussing mutual support and the sharing of resources. In the first half the three researchers on the project, Gweltaz Duval-Guennoc, Sasha Wanasky and Preben Vangberg presented the work they have done so far and in past projects. In the second half, the other researchers and industry professionals in attendance were given the opportunity to present their own work. These included developing corporal conversational agents that are able to hold engaging conversations to be used in the care of the elderly, with hopes to expand the project to Breton in the near future. Another project was presented by Michel Mermet, a volunteer from *Skol Uhel ar Vro*, the cultural institute of Brittany. Their linguistic section recently decided to gather and produce aligned data in order to supplement the existing corpora with material from the Vannetais dialect, and groups 13 volunteers aligning translations and translating material. The translator Lena Catalan Marcos reported on a project she has been working on for several years (*bazh-valan*). The project aimed to pair Breton learners with willing native speakers of the language. The project ultimately did not gain a lot of traction, but it did produce a map which contains contact information of these willing native speakers, and videos from various dialects (hosted online on the website of Bretagne Culture Diversité[22]). The resource can be used for future projects in data collection since these individuals are passionate about the continuation of their language and well-connected in their local communities. They recently tagged the videos as free of use. Finally, the round table discussion also included a presentation of the newly approved *Yezh ar Vro* project, which aims to address the lack of Breton speech data and the lack of audio resources for adult Breton classes at the same time. The project includes the creation of a platform facilitating collaborative transcriptions with an ASR among them.

Apart from presenting the work of the participants, the round table discussion also addressed other issues around licensing, collaboration between researchers and the industry and the exportability of the developed tools and resources for other languages.

Overall, the discussions allowed each participant to raise issues they have encountered and receive advice or support from the other attendees. This shows that the establishment of this network can allow researchers, industry professionals and volunteers to share resources and develop strategies to effectively create tools and technologies for the minoritized languages.

[22] https://www.bcd.bzh/bazhvalan/br/bazhvalan/

5.2.4 Media Coverage of Breton Visits

On the last day of the Breton visit, our Breton partners had arranged a press conference for local news organizations, the outcomes of which were items on the Ouest-France French language news website[23] and a report on the France 3 Bretagne television news[24].

6 Future Direction of Work

The network intends to continue working closely together on language centric AI that mutually benefit our bilingual communities in as many economic and cultural contexts. In many discussions online and face-to-face during visits it became obvious that a very wide scope of technological work exists that both languages could address in a joint manner. We were already able to identify a number of future possible funding possibilities and have identified more areas of technological co-operation since the world of AI is evolving at an incredible pace. Some possible developments are natural progressions from current work, but it is also good to take a more strategic and long-term view of what is needed.

6.1 Extending Current Activities

6.1.1 Speech Recognition

Work on Breton and Welsh speech recognition will continue independently in our respective projects. Wales has benefited significantly from the Breton expertise and we now have an additional dimension to our research and provision to help Welsh companies and organizations. But much work needs to be done in optimizing the models so that all the requirements and needs of our communities are met.

6.1.1.1 Improved Training Environment

The portable and reproducible training environment based on Docker that was developed in this project still has a number of areas for improvement that could be worked on together. Aspects that can be improved in future:

1. Train language models with the OSCAR text corpus[25]. OSCAR is a widely used multilingual text corpus that has been built from web crawls. The Welsh and Breton subsets contain texts millions of words in size. A larger text training set might improve the accuracy of transcribing.
2. Kaldi-Vosk uses the SRI Language Modelling Toolkit[26] for text generation. While its source code is openly available, its terms and conditions[27] however do not allow for commercial

[23] https://www.ouest-france.fr/high-tech/intelligence-artificielle/gallois-et-bretons-ont-travaille-de-concert-pour-creer-une-intelligence-artificielle-de-traduction-77fc424c-0673-11f0-b36c-40f949dc6455

[24] https://www.youtube.com/watch?v=NP83nwOpjD0

[25] https://oscar-project.org/

[26] http://www.speech.sri.com/projects/srilm/

[27] http://www.speech.sri.com/projects/srilm/docs/License

activity without further licensing. Using alternative more permissively licensed language modelling toolkits should be explored[28]

3. Our speech recognition models will require training with added noise augmentation. This aims to improve their robustness and performance in real-world noisy environments.
4. We would like to actively address potential biases in our speech recognition models. This will involve careful data curation, targeted evaluation metrics, and potentially the implementation of bias mitigation techniques during training to ensure fairer and more equitable performance across dialects and demographics.

6.1.1.2 Multi-Style Transcription

A common requirement that could be a focus of further collaboration is that of the style of transcribed text. That is, does the user require a very verbatim transcription or a text with normalization and corrections that are better suited for subtitles and easier reading. Work in other languages and universities (Poncelet & Van Hamme) are investigating this and recent papers suggest that "multi-style" speech recognition models are, in the opinion of network member Associate Loïc Grobol, attainable for Welsh and Breton, given the data and resources at our mutual disposal. The research would also improve our methods for constructing training datasets for multi-styled transcription of spontaneous speech and be a very exciting opportunity to conduct advanced research on training speech models on high performance AI computing clusters in France (and Wales).

6.1.1.3 Pre-transcriptions Website for Linguists

In the meantime, in support of the Yezh ar Vro project during its ramp up, and in support of ongoing collaboration and sharing of experiences on speech data collection, Bangor University will continue to host the Anaouder website that facilitates obtaining pre-transcriptions of Breton speech audio. Bangor will support whatever decision the Yezh ar Vro project makes in the mid-term, for example if they subsequently decide to host, specialize or deploy an alternative with more features or even decide to utilise an alternative solution. In fact, the software is only intended as a temporary solution, but it could have the potential for continued collaborative software development for utilization in other project contexts.

6.1.1.4 Anaouder Graphical Transcription Software

Network member Gweltaz Duval-Guennoc is actively developing an AI centric transcription software, aiming to accelerate the transcription efforts and lower the costs necessary to build ASR datasets for Breton, as well as subtitles in this language. This is a joint project with Dizale, a Breton dubbing organization in Kemper, which received a grant from Region Bretagne. The software, aimed at end users, is due in September 2025 under an Open-Source licence, and as such could be easily adapted to work with Welsh (using the newly trained Welsh Vosk model) and other languages.

Members from Cymen confirmed the use they could have of such a tool, and potential gains over currently used software (specifically: simplicity, and compatibility with macOS).

[28] https://alphacephei.com/vosk/lm (see Language Modelling Toolkits)

Figure 4 - Preview of Anaouder annotation software, including the Breton Vosk model.

6.1.1.5 Fine Tuning and Adaptation

Network members will continue to fine tune popular pre-trained models such as OpenAI's Whisper and Meta AI's wav2vec2. For both types of fine tuning, large collections of transcribed speech is required. Our joint work however informed us of how Kaldi-Vosk based models could be improved, in particular for named entity recognition, with only textual resources. This suggests that Kaldi-Vosk adaptation may be easier, more convenient and more effective.

In future collaborative work, we could extend our experiments and conduct a more rigorous verification process that could impact on democratizing further speech recognition and/or product/service development where the barriers to domain (or customer) specific speech recognition are significantly reduced.

6.1.1.6 ANR Project YAR

The network will be involved and/or kept informed about the progress of the Yezh Ar Vro project.

It is a data collection and tool development project funded by the ANR (the French national research agency), it hinges on the idea of leveraging academia/education/archives collaborations to crowdsource field recordings, transcriptions, and ASR models for Breton (although it is in principle generic enough that it could apply to any language). The proposed backbone workflow is:

- Using a mobile application (yar.app) developed for the project, language learners record native speakers in their community (family, neighbours, friends) in short interviews and upload their recording to a web platform (yar.bzh).
- Back in the classroom, the recordings are checked by teachers for quality and appropriateness, then transcribed collaboratively by learners and teachers proof check the transcriptions.
- Checked transcriptions can then be used to generate more practice material like fill-in-the-blank transcription etc.

Data (recordings and transcriptions) is shared publicly on the web platform (data privacy and ownership is managed by asking participants to agree to this use of the data they produce).

Apart this main data production loop, other efforts contribute to grow the data collection:

- Existing sound archives are contributed and transcribed by professionals and volunteers from the Dastum association.
- Volunteers can contribute and proof check transcriptions via yar.bzh
- Open source ASR models are developed from the data gathered in the project and can be used to pre-transcribe the recordings, providing a different, possibly easier practice for students

In the end, the project helps the integration of language learners in the local linguistic communities, provides speech understanding teaching material, and speech recognition datasets and models.

6.1.2 Text-to-Speech

During his visit to Bangor, Breton network member Gweltaz Duval-Guennoc found Bangor's work on Welsh text-to-speech to be very interesting and another potential area of speech technology for collaboration.

Text-to-speech voices and data had already been developed for Breton by the IRISA lab in Lannion (Guennec & al. 2022[29]), distributed by the Ofis Publik Ar Brezhoneg, but the models are restricted and not generally available, thus of limited use to developers and commercial use. Parts of the training data however were available, and thus a Breton open source TTS model was trained in Bangor using scripts and GPUs originally intended for Welsh.

Open source projects for both languages have therefore been established by the network members that could serve as an additional focus for future work and exchange of expertise.

[29] Guennec, David, Hassan Hajipoor, Gwénolé Lecorvé, Pascal Lintanf, Damien Lolive, Antoine Perquin, Gaëlle Vidal. 2022. 'BreizhCorpus: a Large Breton Language Speech Corpus and its use for Text-to-Speech Synthesis', *The Speaker and Language Recognition Workshop* (Odyssey 2022), 263-270, text.

6.1.3 Large Language Models

6.1.3.1 Breton and Welsh Chat Models

During visits, anecdotal comparisons of Breton and Welsh support in LLMs such as ChatGPT, Claude were made as well as sharing information on work on improving open source models in Bangor and Brittany for Welsh and Breton. Projects such as GweLLM[30] and Bangor's first initial fine tuned LLM model[31] were presented to each other.

It was obvious that there is very active work ongoing and that the network should attempt to increase collaborative efforts in this field.

6.1.3.2 Speech Large Language Models

Multi-modal large language models for Breton and Welsh would be another area of future collaborative investigation. Much research has started recently on the construction and benefits of so-called 'Speech LLMs' [32] [33], that is, a single LLM model (such as LLaMA) that accepts and generates speech and responds with very low latency. This enables real-time, full-duplex and simultaneous responses (such as translations) while speaking.

This would replace our current cascade of multiple models that fulfil different tasks. (speech recognition -> machine translation -> text to speech). This does not undermine or negate the need for task specific models, but for use cases such as simultaneous speech to speech translation, a single multi-modal language model is an essential next step for our language communities. Open-source models are already appearing that are providing these capabilities for the larger languages. We would enjoy exploring, replicating or fine-tuning these models together for the needs of bilingualism in Wales and Brittany. An example candidate 'Speech LLM' model (from France) is called Hibiki[34] which translates French speech to English simultaneously and in real-time. If we were really ambitious, then one model could translate between all four languages.

6.1.4 Applications

Our communities are in more urgent need for applications that dictate our research priorities. In other words, all research in Welsh and Breton language centric AI has to have an immediate or obvious practical element or justification. Models trained for a particular task are not useful unless they are relevant and available to speakers in everyday software.

A common theme throughout our discussions on applications was the need for locally hosted models that run on desktop, laptops and even handheld devices such as phones and tablets.

[30] https://github.com/blackccpie/GweLLM
[31] https://huggingface.co/techiaith/llama-3-8b-instruct-ctp-cy
[32] https://arxiv.org/abs/2412.18495
[33] https://arxiv.org/abs/2409.06666
[34] https://github.com/kyutai-labs/hibiki

6.1.4.1 Speech Enabled Keyboards

One particular use case or application that the network identified as being a priority were for speech models integrated into keyboards on mobile devices, be it Android or iOS. Support for adding custom speech models into an OS, such as Android and iOS, is very fragmented. Some opportunities do exist that could be explored jointly in the future, e.g. adding Kaldi-Vosk ASR into your Android phone's keyboard, are more possible than others[35] and could be a viable and achievable prototype in any continued further work.

6.1.4.2 Voice Assistant

Bangor has provided a Welsh language voice assistant app, called Macsen[36], similar to Alexa, for Android, iOS and online from 2020. All the key language technology models employed by the app are however cloud based. Future work and collaboration could improve Macsen with offline speech recognition and text to speech models as well as extend to multilingual support for four languages (Welsh, English, Breton and French).

6.2 Strategic and Long-term Considerations

6.2.1 Data

More data is a basic requisite for developing language centric AIs. Data may take the form of text or speech, or may be multimodal, and ways to gather or create more data without breaking copyright laws or violating ethical standards is currently a core issue for the development of AI for all languages. It is of special importance to minoritized languages such as Welsh and Breton who are comparatively poor in suitable digital data in any case. To date, collection and curation of data has concentrated on what was achievable in constrained circumstances, resulting in uneven coverage and issues of quality as well as quantity. A focus for future work would emphasize not only on fulfilling quantity requirements, but also quality so that any gaps and biases are identified and mitigated.

Safeguarding data obtained in long-term secure repositories is of equal importance. Short term projects may lose data at project end, and the fragmented nature of some work may make it difficult to find existing resources. Publishing data in international, stable repositories such as GitHub and the European Language Grid helps, but legacy data, non-digital data and other forms of data not compatible with these also need a permanent home. One such possibility is the National Library of Wales, whose charter extends its remit to all the Celtic languages, including Breton. However there should be no single point of failure in any effort to safeguard data, and distributed systems and mirror sites are seen as a robust way of ensuring longevity and security.

Issues of quality, sustainability and retrieval also point to the need for development of common standards for indexing, annotation, and tagging. Developing common standards hand in hand with any new data curation efforts would help achieve clarity and save time and effort in future projects.

[35] e.g. https://github.com/alphacep/vosk-android-demo
[36] https://macsen.techiaith.cymru/

6.2.2 Empirical Scope

The Welsh-Breton Linguistic AI network has so far concentrated on Welsh and Breton as the languages in which it is mainly interested. However, Cornish could be included as a third closely related Celtic language which is unlikely to be included in any other similar network, where there are already synergies of interest as well as personal connections amongst network members. For some projects, it might be worth reaching out to the other Celtic languages (Irish, Scottish Gealic and Manx), without losing the tight focus a smaller network gives us.

Of the Celtic languages, only Welsh and Irish were included in the linguistic descriptions for digital purposes published in the original METANET 2012 White Paper Series[37], revised and updated in 2023 in the ELE Language Reports[38]. Such descriptions are basic tools for developers, and it is ironic that it is the weakest languages with least resources and most need of accessible digital descriptions are the ones excluded from these reports. In order to advance the case of Breton linguistic AI, and that of linguistic AI in other Celtic and less-resourced languages, comparable reports in Breton and the other languages are needed as a priority.

So far, research and applications for Welsh and Breton have mainly concentrated on the standard language. However, dialects (understood as geographical varieties of language) and other language varieties (for example informal speech and slang) also need attention. Further dialogue and collaboration with projects explicitly working with non standard language is a priority. For example, network members Mélanie Jouitteau and Loïc Grobol are external research partners on the IA Gwened project being conducted by the Cultural Institute of Brittany (Skol-Uhel ar Vro) with the main goal of building a parallel Breton-French corpora that includes dialectal variations, particularly the Vannetais dialect, which remains underrepresented online and lacks dedicated digital tools.

Use cases such as automatic subtitling of television drama or transcription of oral stories can not be effective unless they can capture how people actually speak, therefore enabling the capture, analysis and efficient annotation of different varieties needs to be prioritized.

6.2.3 Economic Dimensions

Visits to M-Sparc Science Park and local companies showed the economic potential of linguistic AI for Welsh and Breton. Rather than being an economic burden, investment in research and development in this field increases the profitability of local businesses in peripheral economic regions in Wales and Brittany. The technologies and approaches exploited are cutting edge and serve well in equipping developers and businesses in both communities with valuable AI skills.

Sectors that stand to benefit greatly from linguistic AI in Welsh and Breton are the creative sectors (television, video, multimedia, music, etc), software and games development, health, second language learning, as well as the translation industry. Developing bilingual provision for the domestic market, be it with English or French, helps companies improve their understanding of multilingual

[37] https://european-language-equality.eu/meta-net-white-paper-series/
[38] https://european-language-equality.eu/ele-book/

applications and software, thus preparing them for expansion into multilingual markets if they so wish.

It is imperative therefore that project outputs where public money has been invested are released under permissive open source licences. Appropriate consideration should be given to issues of IP ownership at the beginning of projects, having due regard of course to the rights of individuals and linguistic communities, and any matters of confidentiality and security. Further work is needed to clarify IP and ownership issues in the difficult area of linguistic AI, but the presumption should always be that what was paid for with public money should be publicly available for use by all researchers and developers.

6.2.4 Funding

Another year of Agile Cymru funding was recently announced by the Welsh Government, which the network will examine as early as possible in the new UK financial year with a view to making an application on further technological pilots and cross-region visits. The Breton partners have taken contact with the Brittany Region in order to unlock their B-Monde grant with some local partners. In the meantime, as stated in the first application for setting up the network, Bangor will continue working on a UK Research Council grant application for greater collaboration on Celtic language technologies. Since the European Union also has a role in supporting its languages we will maintain a watching brief on what opportunities derive from the European Language Equality initiative and the Horizon Europe programme.

While small short-term grants are welcome as a way of piloting new ideas and filling some gaps in provision, there is a need for longer term, coherent projects that have the stability and room for innovation provided with longer-term investment. Cross-region projects also enable better pooling of resources, and offer cost-effective solutions to the development of common tools and applications for Welsh and Breton.

Of special interest is the Alliance for Language Technologies European Digital Infrastructure Consortium (ALT-EDIC) created by the European Commission in February 2024. According to its website[39] its activities *"will lead to enhanced artificial intelligence (AI) solutions capable of understanding and generating language taking into account Europe's linguistic diversity and cultural richness."* Its mission focuses on 5 core activities that align well with the ambitions of our Welsh-Breton network for linguistic AI development including long term development and curation of data, and language models, together with evaluation and wider outreach to the public and language communities. ALT-EDIC is poised to begin its operations in Spring 2025 and is likely thereafter to announce some seed funding opportunities. However, Wales, as part of the UK, is not part of the member states or associated countries partaking in this mechanism and is therefore at present excluded from it. Brittany of course might have other opportunities for participation. However, we believe that Wales-Brittany cooperation is valuable in its own right, and will seek other avenues of collaboration, including joining seeking grant opportunities. We will therefore continue to seek joint funding opportunities for our common aims.

[39] https://alt-edic.eu/about-us/

7 Reflections and Future Work

Amongst knowledge exchanges already identified were opportunities for Welsh researchers to take advantage of tools currently being developed for Breton for real-time speech recognition and support for prototyping a live transcription model for Welsh. Welsh researchers learnt of Breton projects on collecting and preparing further data for speech based AIs, thus identifying common practice and synergies, as well as discrepancies, especially with regard to geolocation metadata as well as tagging different accents and dialects.

Breton researchers have learnt about current developments in Wales regarding fine tuning and/or domain adaptation of speech recognition models, thus improving accuracy in particular with regard to ensuring the correct transcription of the names of people, places and companies (i.e. named entities).

Working collaboratively between January and March 2025, experimental developments included adapting a live transcription model for Welsh, designing an AI named entities model for both Welsh and Breton, and evaluating integrating speech and machine translation together. Joint workshops were held in Wales on 20-21 February, and in Brittany on 20-21 March, including Round Table discussions in order to gain a wider understanding of the needs and aspirations of various stakeholders: researchers, industry and language communities. These discussed such questions as the priorities for developing conversational AI for Welsh and Breton, how to upskill your workforce to meet the new challenges, how to create sustainable resources, how to take advantages of new and evolving methodologies, and how to make the best use of data collection methods such as crowdsourcing for less-resourced languages.

Project discussions and results have been gathered together to create this report in the form of a concise Roadmap for all stakeholders to guide us through what is at times a confusing and rapidly changing landscape and enable us to plan future collaborations and build upon the progress already made in this new area of Welsh-Breton cooperation.