Institution:
Bangor University, 10007857

Unit of Assessment:
UoA 26 - Modern Languages and Linguistics

Title of case study:
Language technology for Welsh as a less-resourced language

Period when the underpinning research was undertaken:
2001 to 31 December 2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
1) Delyth Prys	1) Professor of Language Technologies	1) September 2001 - present
2) Dewi Bryn Jones	2) Senior Software Developer	2) October 2002 - present
3) Gruffudd Prys	3) Senior Terminologist	3) August 2001 - present
4) Dr Sarah Cooper	4) Lecturer, School of Languages, Literatures, Linguistics and Media	4) February 2014 - present
5) Dr Myfyr Prys	5) SMART Partnership Contact/KTP Associate	5) September 2014 - December 2020

Period when the claimed impact occurred:
1 August 2013 – 31 December 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact

Research undertaken by Bangor University's Language Technologies Unit (LTU) in the field of language technology for Welsh as a less-resourced language has enabled the development of on-line resources (proofing tools, dictionary apps, text to speech, speech recognition and machine translation), transforming the use of Welsh in digital environments and impacting daily lives across Wales. Research-led development of resources has helped to promote Welsh as a modern, forward-looking language, as well as enhancing practice in industry and informing government policy. Impact has extended beyond Wales, influencing the digital language equality roadmap internationally and informing the practice of global concerns.

2. Underpinning research

Building on longstanding expertise in the field of language technologies and national terminologies in the context of less-resourced languages **[3.1]**, Bangor's Language Technology Unit (LTU) research has focussed on identifying the needs of minority languages such as Welsh that lack the large electronic resources of major languages **[3.2]** and fostering the development of methodologies which enable the production of the requisite digital tools **[3.a]**. Research and development have centred on two main areas: ***Lexica and Grammar*** and ***Text-to-Speech / Speech Synthesis.***

Lexica and Grammar
By mining bilingual databases from LTU terminology standardization and dictionary activities, researchers were able to develop lexica, part of speech taggers and morphological descriptions of Welsh. *Cysgliad*, a comprehensive Welsh proofing tool, comprises the *Cysill* spelling and grammar checker and the *Cysgeir* dictionary compendium. Cysill is based on the former *CySill* program (1995). Its code base and content were revamped and expanded in 2004. Researchers also expanded its Welsh lexicon and rule-based grammar checker of over 300 general grammar rules and over 200 mutation rules, rewriting its tokenizer, segmenter, morphological analyser, lexicon, part of speech tagger and lemmatizer. A free on-line version (*Cysill Ar-lein*) launched in 2009 allowed the automatic collection of an ever-increasing text corpus, currently over

300,000,000 words, which researchers use to produce language models, frequency word-lists and speech prompts **[3.3]**. *Cysgeir* contains eleven bilingual dictionaries also used to mine lexical and morphological information and run the Vocab word-by-word translator **[3.4]**. Researchers also developed *Maes-T* as a bilingual master database enabling online dictionary development and extraction to different speech and text products such as the *Ap Geiriaduron*, *Y Termiadur Addysg* and *Geiriadur Termau'r Coleg Cymraeg Cenedlaethol*. In addition, researchers adapted *Maes-T* to publish an online version of the *Welsh Academy Dictionary* and the *Gerlyver Kerenewek* online English/Cornish dictionary **[3.5, 3.b]**.

Text-to-Speech / Speech Synthesis tools

Research undertaken as part of the *Welsh and Irish Speech Processing Resources (WISPR)* project (2004 - 2007), in collaboration with three Irish Universities who developed similar Irish language resources, enabled the development of new and improved letter-to-sound rules for Welsh, recording scripts, and speech data recorded by voice talents. This resulted in two Welsh Windows-based synthetic voices for end-users, and Welsh text-to-speech resources for other developers to build new voices. Researchers adopted newer neural network methods of developing speech synthesis and speech recognition by gathering and analysing substantial datasets. In 2014, an innovative *Paldaruo* app was developed to crowdsource speech recordings as the *Paldaruo Speech Corpus* using smartphones' own mikes and speakers rather than traditional recording equipment **[3.6]**. Researchers have also worked in partnership with Mozilla on their multilingual Common Voice crowdsourcing project since 2016. By August 2020 the Welsh speech database developed by LTU contained over 1,285 crowdsourced Welsh voices and 85 hours of recordings. Researchers used this data to develop 'Macsen', the first Welsh language digital personal assistant, and 'Trawsgrifiwr' the first Welsh transcriber.

In 2015, LTU launched the *Welsh National Language Technologies Portal*. Designed to give easy access to their tools and resources (over 45 in number), it links to international repositories (e.g. GitHub and Metashare) where code and data are stored **[3.7, 3.c - 3.f]**. Experiments with Machine Translation (MT), reusing our lexical tools and resources, led to new Welsh/English MT engines **[3.8, 3.g, 3.i]** used in the translation industry.

In April 2020 Welsh Government asked LTU to develop and publish a free downloadable version of Cysgliad to help with remote education and working during the COVID-19 pandemic **[3.i]**. In addition, the research group were funded in October 2020 to help mitigate home working during the COVID-19 pandemic by developing a multilingual conferencing system with machine translation based on LTU research **[3.j]**.

3. References to the research

Research Outputs

3.1 **Prys, D**. (2011) Developing National Terminology Policies: A Case Study from Wales. *Journal of Hungarian Terminology, **4**(2), 160-168*. DOI (Peer-reviewed journal article).

3.2 **Prys D**., **Prys G**. and **Jones D. B**. (2015) Quantifying the Use of Digital Welsh-language Language Resources. *7th Language & Technology Conference*, Poznan, Poland, 27-29 November 2015. (A copy available on request).

3.3 **Prys D**., **Prys G**. and **Jones D. B**. (2016) Cysill Ar-lein: A Corpus of Written Contemporary Welsh Compiled from an On-line Spelling and Grammar Checker. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3261-3264. Link

3.4 **Jones D. B**., **Prys G**. and **Prys D**. (2016) Vocab: a dictionary plugin for websites. *Second Celtic Language Technology Workshop,* 93-99. Link

3.5 **Prys D**. (2020) Adapting a Welsh Terminology Tool to Develop a Cornish Dictionary. *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*: 235-239. Link

3.6 **Cooper S**., **Jones D. B**. and **Prys D**. (2019) Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology. *Information,* **10**(8), 247-259. DOI (Peer-reviewed journal article) Submitted to REF2021 (REF identifier UoA26_17).

3.7 **Prys D**. and **Jones D. B**. (2018) National Language Technologies Portals for LRLs: a Case Study. *Language and Technology Conference: LTC 1015 Human Language Technology. Challenges for Computer Science and Linguistics,* 10930: 420-429. DOI (Peer-reviewed conference proceeding).
3.8 **Prys, M**. and **Jones D. B**. (2019) Embedding English to Welsh MT in a Private Company. *Proceedings of the Celtic Language Technology Workshop*. European Association for Machine Translation. W19-69, 41-47. Link

Terminology & Lexicography Grants
3.a **Prys, D.** (2011 – 2020) *Termiadur Addysg*. Welsh Government. GBP1,042,958 (Bangor University: R59G02)
3.b **Prys, D.** (2017 – 2019) *Cornish Dictionary Development*. Cornwall Council. GBP22,000 (Bangor University: R59G14, R59G17)
Speech and Language Technology Grants
3.c **Prys, D.** (2014 – 2016) *Technoleg a Chyfryngau Digidol / Technology and Digital Media*. Welsh Government and S4C. GBP49,779 (Bangor University: R59G08)
3.d **Prys, D.** (2016 – 2020) *Macsen*. Welsh Government. GBP239,997 (Bangor University: R59G10)
3.e **Prys, D.** (2013 – 2014) *GALLU: Gwaith Adnabod Lleferydd Uwch / Advanced Speech Recognition for Welsh*. Welsh Government. GBP77,703 (Bangor University: R59G07)
3.f **Prys, D** (2017 - 2018) *Lleisiwr*. Welsh Government. GBP20,000 (Bangor University: R59G13)

Machine Translation Grants
3.g **Prys, D.** (2017 - 2019) *Knowledge Transfer Partnership* (KTP with Cymen Cyf. Innovate UK. GBP130,727 (Bangor University: R59K01)
3.h **Prys, D.** (2019 – 2020) *SMART Partnership with Cymen Cyf*. Welsh Government. GBP41,952 (Bangor University: R59K02)
3.i **Prys, D.** (2020 - 2021) *Technology and the Welsh Language*. Welsh Government. GBP347,950 (Bangor University: R59G18)
3.j **Prys, D.** (2020 - 2021) *ONCON Online Conference Platform*. Innovate UK. GBP23,577 (Bangor University: R59G19)

4. Details of the impact

Research undertaken by Bangor's Language Technology Unit (LTU) has substantially modernized approaches to the use of the Welsh language and informed practice more broadly in the area of less-resourced languages. Research has had impact in four key areas: ***Influencing Policy***; ***Fostering the Daily Use of Welsh***; ***Developing Assistive Technology***; and ***Supporting Industry***.

Influencing Policy
Researchers have influenced policy in Wales and globally. Professor Delyth Prys and Dewi Jones have served as advisors on the Welsh Government (WG) Welsh Language Technology Board since 2012, helping inform the approach to language revitalization through language technologies in WG's strategy *Cymraeg 2050* (2017). The subsequent *Welsh Language Technology Action Plan* (2018) references LTU work, highlighting speech technology, machine translation and conversational Artificial Intelligence (AI), together with terminology, lexicography and corpora resources **[5.1]**. On a global level, D. Prys, Jones, Andrews and G. Prys have advised on UNESCO's LT4All initiative, the Digital Language Diversity Project (2015-2018) **[5.2]** and the successful EU Parliament Resolution on Language Equality in the Digital Age (adopted 11 September 2018). The Rapporteur, an MEP, referenced LTU research in her parliamentary speech **[5.3]**. This resolution has now resulted in the PPPA-LANGEQ-2020 call for 'Developing a strategic research, innovation and implementation agenda and a roadmap for achieving full digital language equality in Europe by 2030.'

Fostering the Daily Use of Welsh
Research has directly influenced the public's use of Welsh. *Cysgliad* is credited by the Welsh Language Commissioner with helping members of the public gain confidence in writing Welsh, as

well as being an essential aid in education and for translators and professional administrators **[5.4]**. Language technology tools developed by LTU researchers are in daily use across Wales The Welsh Language Commissioner mandates all public authorities to install Welsh spelling and grammar checkers on their computers for which *Cysgliad* is the only available product **[5.5]**. During October 2020, approximately 1,000,000 words per week were spell- and grammar-checked on the free website *Cysill Ar-lein*. Between September 2015 and October 2020 approximately 270,000,000 words were added to the *Cysill Ar-lein* corpus. As of 7 October 2020, *Cysgliad* is licenced to 492 public and private authorities in Wales, and individual licences (paid for) total 5,392. In 2020, the Welsh Government sponsored free *Cysgliad* licences for the public, schools and SMEs to help during the COVID-19 pandemic. During the first six months of the offer (between May and October 2020), an additional free 4125 licences were downloaded.

Between August 2013 and October 2020, the *Ap Geiriaduron*, including *Geiriadur Termau'r Coleg Cymraeg*, was downloaded 230,646 times. A review (24 July 2019) on Google Play states "*This is an incredible app for searching a word I don't know in a welsh text*". Between March 2015 and October 2020 Vocab, with a unique identifier key downloadable from the National Terminology Portal, was used on 68,537 unique Welsh web pages and on 19 websites, including BBC CymruFyw and Golwg360, and 544,844 word searches were carried out. Development of the Macsen personal assistant has changed attitudes on the possibilities of Welsh language technologies across the generations, evidenced by the positive responses of schoolchildren employing Bangor language technology during coding sessions to make a robot talk in Welsh **[5.6]**.

Developing Assistive Technology

Research in Welsh speech technology has improved the lives of visually impaired Welsh-speakers, as well as those with reduced mobility and communication difficulties **[5.7]**. Users were previously unable to access Welsh text-to-speech, and Welsh voice-activated applications. Tools and resources for Welsh speech technology, published by LTU under open licences, fed into improved synthetic voices developed also by commercial companies e.g. Ivona (since acquired by Amazon). The Trawsgrifiwr program is also being used by the National Library of Wales to help their volunteers transcribe audio material from their broadcast archives for their Voice2Text project. The Royal National Institute of Blind People (RNIB) and others use these synthetic Welsh-speaking voices to enable visually impaired users to access Welsh text-to-speech. In April 2017, the speech recognition research led to *Lleisiwr*, a joint Bangor University / NHS project whereby users about to lose their power of speech can bank their voice and have it rebuilt as a synthetic voice, as shown in an S4C documentary programme 'Drych: Achub Llais John Wyn' (April 2019) **[5.8]**. Researchers were awarded the Royal College of Speech and Language Therapists' Giving Voice Award for their "*outstanding commitment to raising awareness of people's speech, language and communication needs at a national level*" in September 2019 **[5.9]**.

Supporting Industry

Research has helped companies locally and globally to develop new products, providing economic stimulus to peripheral areas and giving Welsh an international presence. In 2017 Facebook expressed appreciation of LTU-developed resources: "*thanks a lot for the pointer […] since this is Apache 2.0, we have been able to download this data already*" **[5.10]**. Pioneering use of crowdsourcing for the *Paldaruo* speech-corpus has informed methodologies used by major international companies to gather speech data for multiple languages. Mozilla adopted this approach for its Common Voice project **[5.11]** and the *Paldaruo* corpus itself has been downloaded 73 times by other developers. The containerized version of the Welsh/English machine translation Moses system in Docker enables translation companies to build their own translation engines from legacy translations. It has been downloaded 415 times through the National Portal. It led to a Knowledge Transfer Partnership (KTP) between 2017 and 2019 **[3.g]** with a local translation company, Cymen, (awarded an A [outstanding] grade) and also led to funding from the Innovate UK Accelerated COVID-19 Response Call **[3.j]** in 2020, enabling LTU to work with two local companies, Animated Technologies Ltd and Zero Dependencies Ltd, and Menai Science Park (MSparc), to develop Inclusive and Advanced Online Conferencing, including multilingual translation functions.

5. Sources to corroborate the impact

5.1 **Language Technology Action Plan Welsh Government (2018)**. References LTU work, highlighting speech technology, machine translation and conversational Artificial Intelligence (AI), together with terminology, lexicography and corpora resources.
https://gov.wales/sites/default/files/publications/2018-12/welsh-language-technology-and-digital-media-action-plan.pdf
5.2 **Testimonial letter from Digital Language Diversity Project Coordinator, National Research Council of Italy** (Reporter on the impact process). Testifies to the influence of Bangor's research on European research agenda for minority and less-resourced languages.
5.3 **Testimonial letter from MEP Rapporteur for the Culture and Education Committee of the European Parliament Language Equality Resolution 2018/2018 (INI)** (Participant in the impact process). Testifies to Bangor's critical role in developing and taking forward the roadmap for achieving full digital language equality in Europe by 2030. Prys et al acted as advisors for MEP as Rapporteur for the successful European Parliament Resolution on Language Equality in the Digital Age leading to the European Commission's call for a strategic roadmap and agenda LANGEQ2020.
5.4 **Testimonial letter from the Welsh Language Commissioner (WLC)** (Participant in the impact process). Corroborates the influence of LTU tools and resources on the general public's use of Welsh and the development of bilingual policies by public bodies and industry.
5.5. **The Welsh Language Standards (No. 1) Regulations 2015, No. 996 (W. 68) p.48**. Welsh Language Commissioner standard 120 clearly states that staff must be provided with computer software for checking spelling and grammar in Welsh, and provide Welsh language interfaces for software (where an interface exists) in the workplace; Bangor's Cysgliad is the only available software that can fulfil this statutory requirement.
http://www.legislation.gov.uk/wsi/2015/996/pdfs/wsi_20150996_mi.pdf
5.6 **Techiaith Vimeo video (2015)** Gwers codio Cymraeg ar y Raspberry Pi yn Ysgol Garndolbenmaen (*Welsh language Raspberry Pi coding lesson in Garndolbenmaen Primary School*). Evidences the positive responses of schoolchildren employing Bangor language technology during coding sessions to make a robot talk in Welsh. (Language: Welsh) (Evidence submitted is a screenshot).
https://vimeo.com/121462473
5.7 **Testimonial supporting statement from the Head of Project 2050, Welsh Language Division, Welsh Government** (Reporter and participant in the impact) on behalf of the Welsh Government Minister for International Relations and the Welsh Language. Corroborates the impact on public use of Welsh, development of Welsh language assistive technology, on government policy (Language: Welsh, translation available on request).
The importance of Bangor's research team in developing the technology is further supported by **the Minister for International Relations and the Welsh Language in a Senedd committee, referring directly to Bangor having the only relevant expertise.** (Evidence submitted is a screenshot).
http://senedd.tv/Meeting/Clip/a65b2aac-e0c1-4fdc-b7a7-4ef83507c433?inPoint=00:54:50&outPoint=00:56:36
5.8 **North Wales Daily Post news article** of individual with throat cancer using LTU Welsh text-to-speech voice which clearly credits Bangor University for this innovative technology.
https://www.dailypost.co.uk/news/north-wales-news/welsh-speaking-voice-box-allows-16144287
5.9 **Testimonial letter from the CEO, Royal College of Speech and Language Therapists** (Participant in the impact process). Confirmation of award for LTU's speech technology research and impact.
5.10 **Unsolicited email (2017) from Facebook** expressing appreciation of LTU-developed resources (Reporter on the impact). Corroborates the global reach of the research; supporting companies to develop new products and giving Welsh an international presence.
5.11 **Testimonial supporting letter from Manager and Technical Lead of Mozilla's Machine Learning Group** (Participant in the impact process). Corroborates details of LTU's partnership with Mozilla and influence on multilingual development of Common Voice.