Gathering Data for Speech Technology in the Welsh Language: A Case Study

Delyth Prys, Dewi Bryn Jones
Language Technologies Unit, Bangor University
Bangor, Wales, UK
{d.prys, d.b.jones}@bangor.ac.uk

Abstract

Less-resourced languages face additional challenges in the creation of tools and resources for speech recognition applications. These include lack of funding, sparsity of data and shortage of experts with relevant skills. On the other hand there are also opportunities to be had from tapping into committed communities of language activists, and potentially developing innovative solutions to common problems that may be applied elsewhere. This paper describes a recent series of short-term projects for the Welsh language that have used crowdsourcing methodologies, together with data from Wicipedia (the Welsh Wikipedia) and existing Welsh corpora, to further advance the field. They have also borrowed and adapted open source tools, such as MaryTTS and Mozilla CommonVoice that were already freely available. In addition this paper provides some pointers towards further needs and solutions for speech technology in less-resourced languages, aiming at a coherent, long-term approach that may be applicable in many environments.

Keywords: speech recognition, less-resourced languages, Welsh

1 Introduction

Automatic speech recognition (ASR) technology is the most critical component in intelligent speech interfaces that are becoming increasingly popular amongst consumers who wish to access web-based information and services. Such platforms have recently become viable due to advances in the use of neural network methods to reduce speech recognition word error rates. However, ASR research has been primarily carried out for English and other major languages, where the vast amounts of text and speech data required for training exist, or where it is commercially viable to collect and deploy it.

It remains challenging however to develop ASR for less-resourced languages with limited or no training data, and where there is no immediate or direct commercial benefit for private sector actors. In such cases, stimulation and investment is required from public organisations or foundations that desire, or are under legal obligation, to support speakers of less-resourced languages.

In 2013, the Welsh Government published its *Welsh Language Technology and Digital Media Action Plan* that provided grants for financing projects aiding any aspect of pairing Welsh language and technology. Due to budgetary and other constraints, these grants could only fund short term projects, with no certainty of continued funding. They needed to produce outputs useful to the public, or to engage the public by other means, and they needed also to show continued innovation, rather than 'more of the same'. In addition, they had to demonstrate good fit to the Government's Action Plan and the priorities named in it.

ASR development is a long-term investment, requiring research and development by many individuals over a number of years, rather than months. In order to advance the development of Welsh ASR therefore, and in the absence of any other sources of funding, a strategy was devised to incrementally develop, through a series of possible projects, a Welsh language digital assistant that could be exploited to guide the research and development of Welsh ASR from the initial position of no data and no support, to incrementally increasing the amounts of data available, and specifically helping towards a digital personal assistant that would be able to understand and respond to oral commands and questions in Welsh.

1.1 Project 0 (2013 – 2014)

Project or stage 0 of our longer-term objectives began with the study of Welsh letter-to-sound rules and production of pronunciation dictionaries as well as the novel development of iOS and Android apps for crowdsourcing a speech corpus. Up until 2014, the Voxforge.org website had been the only possible platform for collecting transcribed speech for a new language. However, our apps brought our crowdsourcing activity up to date in order to facilitate contributions from more recently and increasingly popular mobile devices. The app, named 'Paldaruo' (Welsh for 'to chatter'), provided a small initial collection of 43, specially crafted for phonetic coverage, prompts that users could record at their leisure. Recording all prompts would take approximately 30 minutes and over 400 users duly complied with providing recordings of their voices. The speech data was successfully used to build acoustic models that successfully recognised commands to move a robot arm connected to a Raspberry Pi using spoken commands in Welsh (Cooper et al, 2014).

1.2 Project 1 (2015-2016)

After the end of Project 0, the Paldaruo app was kept in the app stores, and continued to collect speech data from the Welsh language community.

The increasing amount of speech data was put to use in the next project we conducted, namely to expand the capability of Welsh speech recognition to recognize closed and simple, but still useful, questions regarding time, weather and news. These questions in turn would be included as 'skills' into the first prototype version of a Welsh-language digital assistant named 'Macsen'. Macsen would run on a Raspberry Pi, with the Welsh speech recognition engine implemented using HTK acoustic models within Julius.

Example questions (translated) included: "What is the time?", "What is the weather for today?", "What's the news?", "Play me some music", "Play me some Welsh music" (Jones and Cooper, 2016).

1.3 Project 2 (2016 – 2017)

Project 1 gave us a valuable insight of where challenges remained and to strategize for longer term progress whilst still constrained by short term project goals.

Speech technologies are dependent on speech data consisting of audio aligned with textual annotations at segment, word and/or phonetic levels. The Paldaruo speech data thus collected was used to produce the first Welsh language forced aligner, based on Prosodylab Aligner. This would aid future research in increasing and improving the size and quality of Welsh language speech data.

Project 2 also provided an opportunity to re-base Macsen's speech recognition components on the more widely used and popular Kaldi-ASR. A prototype recipe for Macsen's questions was developed and the resulting engine was hosted online and accessible via an API, not only to Macsen but to other software such as the Kõnele app for Android.

Finally, speech interfaces such as Macsen also require text-to-speech capability. Up until project 2, the Welsh TTS options for 'Macsen' were limited to either an open source, but robotic sounding, Festival diphone voice, or a commercially available naturally sounding voice from Amazon Ivona's SpeechCloud. Project 2 provided an opportunity for Macsen to gain open source and naturally sounding voices by developing tools to simplify building unit selection Welsh voices with MaryTTS.

1.4 Project 3 (2017 – 2018)

The Welsh Government in its most recent strategy for the Welsh Language (Cymraeg 2050, 2017) emphasised again the importance of digital technologies, including investing in Welsh language speech technology, to ensure the vitality of Welsh, and the fulfilment of its ambitious plan to double the number of Welsh speakers by 2050.

As part of this commitment, it funded further work on Macsen, and the publication and open dissemination of resources used in its improvement. This led to the current project, which has as its aim extending Macsen's speech recognition capability to being able to recognise more open-ended questions that a user would typically ask a digital personal assistant in order to gain new knowledge.

Wikipedia is an invaluable source for knowledge, especially for less-resourced languages where other available digital data may be scarce. The Welsh language Wikipedia, called Wicipedia, emphasizes creating original content in Welsh, as well as localizing international content where appropriate. The Welsh Government's Cymraeg 2050: Work programme 2017-21 (2017) specifically names supporting efforts to increase the number of Wicipedia pages as one of its aims for the period to 2021. Wikipedia's regular structure also facilitates finding information in its pages, and extracting relevant material to answer questions.

A module or skill in Macsen, enabled by a more developed Welsh language speech recognition, would thus respond to oral questions such as "Pwy oedd Hywel Dda?" (*"Who was Hywel Dda?"*) by reading the first paragraph of the article on Hywel Dda in Wicipedia. This has entailed crowdsourcing greater amounts of richer sound data, as well as the analysis of Wicipedia content and usage.

2. Formulating Recording Scripts for Questions

In our first project to develop speech recognition for Welsh from scratch, sound recordings of a small set of prompts had been sufficient for simple questions and commands. Our experience with the Paldaruo app provided an insight into the nature of participation and engagement levels when crowdsourcing in a less-resourced language community. It was observed that:

- Welsh language speakers were very willing to contribute recordings of their voices
- Contributors would typically record in one sitting/session
- A majority of contributors did not record all of the prompts
- The number of contributors, so far, (without any dedicated significant funding for marketing) is in the 500-600 range.

These observations guided our decisions as to how a larger and richer set of speech data, required for the larger domain of recognising questions to Wikipedia, should be realised. In contrast to assumptions in larger languages, crowdsourcing from a less-resourced language community requires extracting as much speech information as possible from the limited numbers and length of contributions.

Questions to Wikipedia however represent a large domain and thus careful planning would be required as to how the Paldaruo crowdsourcing capability could be expanded to facilitate successful recognition of questions likely to be asked by Welsh language users.

2.1 Subjects covered in Wicipedia

A question and answering module inside Macsen would be perceived as useful if it would answer questions on the subjects typically asked of it. Thus we would have to ensure that at the very minimum, models and thus speech data for these subjects would be sufficiently captured in any crowdsourcing activities.

The Welsh Wikipedia currently contains over 90,000 articles. These are not usually straight translations from the English, and cover specifically Welsh topics or have a unique Welsh slant on international or more general subjects. Judging what would be the most typically viewed subjects in Wicipedia entailed gaining the assistance of the Wikipedia UK Manager in Wales. The project was kindly furnished with a list of analytics websites, in particular Catscan (https://petscan.wmflabs.org) and Topview (https://tools.wmflabs.org/topviews/?project=cy.wikipedia.org). The summation of top views for each month from July 2016 to July 2017 gave surprising results. For example, the most viewed article in Wicipedia by far was about an international pornographic actress. Closer inspection however showed that she did not have an English language article, therefore English queries in popular search engines were pointing towards the Welsh language article.

Having manually weeded out anomalies such as this, and filtered for any articles deemed to contain inappropriate material for family audiences, an initial list of 1000 articles was filtered down to 646 that had received at least 250 visits. The top twenty articles in the filtered list are given in Table 1.

Article Title	*(Translation)*	Visits	Type (object)	Tense (default)
Cymraeg	*Welsh*	12105	Lang	
Saesneg	*English*	10840	Lang	
Cymru	*Wales*	8909	Place	
Unol Daleithiau America	*United States of America*	8688	Place	
Y Deyrnas Unedig	*United Kingdom*	5842	Place	
Ffrainc	*France*	4902	Place	
Yr Ail Ryfel Byd	*Second World War*	4735	Event	Past
T. Llew Jones	*T. Llew Jones*	4687	Person	Past
Lloegr	*England*	4669	Place	
Yr Almaen	*Germany*	4479	Place	
Wikipedia	*Wikipedia*	4339		
Hedd Wyn	*Hedd Wyn*	4159	Person	Past
Sioned James	*Sioned James*	4141	Person	Past
Rhyngrwyd	*Internet*	4131		
Ewrop	*Europe*	4089	Place	
Lladin	*Latin*	4030	Lang	
Caerdydd	*Cardiff*	4012	Place	
Awstralia	*Australia*	4004	Place	
Wicipedia Cymraeg	*Welsh Wikipedia*	3951		
Cynnwys rhydd	*Free content*	3896		

***Table 1** - Edited List of Most Viewed articles in Wicipedia (07/16 - 07/17)*

2.2 Enlarging the prompts set with questions for improved acoustic modelling

Analysis of the most viewed subjects led to a categorisation on the type of question one would naturally ask, such as "Pwy ydy …?" (*"Who is ….?"*) / "Pwy oedd ….?" (*"Who was….?"*) / "Beth ydy…?" *("What is ….? ")* in order to gain knowledge on that subject.

With additional meta-data tagged in the *Type* and *Tense* columns as demonstrated in Table 1, an initial list of questions was generated for all 646 subjects, examples of which as given in Table 2.

Example Question	(Translation)
Beth ydy Cymraeg?	*What is Welsh?*
Beth ydy Saesneg?	*What is English?*
Beth oedd Yr Ail Ryfel Byd?	*What was the Second World War?*
Pwy oedd T. Llew Jones?	*Who was T. Llew Jones?*
Beth ydy Wikipedia?	*What is Wikipedia?*
Pwy oedd Hedd Wyn?	*Who was Hedd Wyn?*
Pwy oedd Sioned James?	*Who was Sioned James?*
Beth ydy Rhyngrwyd?	*What is the internet?*
Beth ydy Lladin?	*What is Latin?*
Beth ydy Wicipedia Cymraeg?	*What is Welsh Wikipedia?*
Beth ydy cynnwys rhydd?	*What is free content?*

Table 2 – Generated Example Questions

This produced a rather limited repertoire of different questions and ways of asking for information. In order to cross-reference with a wider set of possible question, it was decided not to rely solely on Wicipedia, and to seek other sources for examples of questions in Welsh.

The research team undertaking this work was fortunate to have at its disposal the 100 million-word Cysill Ar-lein Welsh language corpus, collected via the free on-line

provision of its popular Welsh language spelling and grammar checker product Cysill (Prys and Jones 2016a). Over 17,000 examples of questions were extracted by simple regular expression matching for the '?' symbol, and omitting any questions that contained capital letters or numbers, in case such questions included private information such as names and/or phone numbers.

Experience with our Paldaruo crowdsourcing app however showed it to be unrealistic to expect a set of more than 17,000 questions to be recorded by thousands of contributors. If contributors numbered only in their hundreds, although a significant achievement for a small language community, there was a danger that a random selection of prompts from too large a prompt set would not provide sufficient phonetic coverage in the speech data. The size of the prompt set would have to be sufficient for a random selection mechanism to allow contributors, between them, to be able to contribute as much phonetic data as possible.

For this task, we were able to approximate a suitably sized and phonetically balanced prompt set by re-using our Welsh MaryTTS resources. Usually these resources are used in crafting recording scripts; however in this utilization we reduced the number of questions from more than 17000 to 300.

A further reduction was achieved by reducing the number of questions with the common beginning "Beth yw..?" (W*hat is*) and relocating subjects into prompts that listed individual words. Some of the individual words are mutated forms that do not usually occur in standalone words, as they are triggered by other words in a sentence. However, some of them include phonemes that were otherwise rare, and so this was deemed the easiest way of including them in a small prompt set. In all a new collection would consist of 270 prompts examples of which can be seen in Table 3.

A fedrwch chi fy helpu I os gwelwch yn dda?	*Can you help me please?*
Faint o'r gloch mae amser cinio yn gorffen?	*What time does lunch finish?*
Faint o'r gloch fydd y bws nesaf yn mynd heibio?	*What time does the next bus go past?*
Faint mae llaeth yn costio?	*How much does milk cost?*
Wyt ti mewn wythnos nesa o gwbl i arwyddo nhw?	*Are you in next week at all to sign them?*
Beth sydd yn digwydd yn y stori yn eich geiriau eich hun?	*What happens in the story in your own words?*
Beth oedd y ffeithiau mwyaf diddorol i ti a pham?	*What were the most interesting facts for you, and why?*
Ble mae cartref gofal agosaf yr Awurdod Lleol ?	*Where is the nearest Local Authority care home?*
Rhyfel Cartref America, Y Brythoniaid, y Chwyldro Diwydiannol	*American Civil War, The Ancient Britons. The Industrial Revolution*
Y Dirwasgiad Mawr, Yr Oesoedd Canol, Y Rhyfel Oer	*The Great Depression, The Middle Ages, The Cold War*
Bob Dylan, Bryn Fôn, Caryl Parry Jones	*Bob Dylan, Bryn Fôn, Caryl Parry Jones*
Ceri Wyn Jones, Cymraeg, Dafydd Dafis	*Ceri Wyn Jones, Welsh, Dafydd Dafis*
Theuluoedd, Toes, Porthaethwy, Cibwts, Rhaeadr, Lliw, Minoaidd, Nymff	*Families, Dough, Menai Bridge, Kibboutz, Rhayader, Colour, Minnoan, Nymph*
Magdalen, Cewri, Ffeuen, Clwyfau, Puw, Sipswn, Llai, Fronhaul	*Magdalen, Giants, Bean, Wounds, Pugh, Gipsies, Fronhaul*
Soia, Deuawd, Prawf, Rois, Teulu, Byw, Ddaw, Amheus	*Soya, Duet, Test, I gave, Family, Live, Will Come, Doubtful*
Bwdhaeth, Botswana, Gewyn, Heddiw, Ebwy, Lleisiau	*Buddhism, Botswana, Ligament, Today, Ebbw, Storyteller, Llangeitho, Voices*
Caerhun, Llew, Arwyllsiad, Ieithoedd, Ehangdir, Ceulan, Bontddu, Nhrwyn	*Caerhun, Lion, Discharge, Languages, Expanse, Hollow Bank, Bontddu, Nose.*

Table 3 - Example prompts with English translations

3. Crowd sourcing voice recordings

When no speech data is available, either due to none actually existing or not suitably licensed, for developing speech recognition for any language, crowdsourcing can be an effective strategy for bootstrapping initiatives. Up until November 2017, the fourth release of the Paldaruo speech corpus had collected 38 hours of speech from 536 contributors which were available according to a CC-BY license. Challenges remain in expanding the corpus, attracting more contributors and in ascertaining its quality for speech recognition development.

A recent welcomed international development regarding crowdsourcing speech data has been Mozilla's CommonVoice project (Mozilla n.d). CommonVoice aims to crowdsource large and publicly available voice datasets in order to foster innovation and healthy commercial competition in machine learning based speech technology. Launched in summer 2017, by January 2018 CommonVoice had crowdsourced, and verified, 254 hours of English language speech, provided by nearly 20,000 volunteers worldwide (The Mozilla Blog, 2017). Its ambition, with the help of open source communities, is to increase the number of hours to the thousands and to build large public-domain datasets for as many languages as possible in the world.

As our case study demonstrates, this ambition will be challenging for languages that have less human and digital

resources, where crowdsourcing dynamics do not exist or cannot scale down effectively.

CommonVoice has allowed us to fork the website code to provide a web-based version of the Paldaruo app. It also has additional features such as the ability to crowdsource verifying the correctness of other recordings. The lack of an easy mechanism for verification and quality control had been a weakness in our previous Paldaruo app. By now the entire corpus is being verified by volunteers and future release of the Paldaruo corpus will similarly claim to have verified hours of speech.

We also embraced the CommonVoice software in order to expand the Paldaruo corpus via a website and to understand how Mozilla's approaches could be scale-down and made impactful for lesser resourced languages. Figure 1 shows a screenshot of our offering at http://paldaruo.techiaith.cymru, with Welsh language text. This translates as "What is speech recognition and why Paldaruo?" with a short explanation on its increasing importance and its ubiquity e.g. in personal devices.

Figure 1 - Screenshot of CommonVoice website at http://paldaruo.techiaith.cymru to crowdsource Paldaruo

Required alterations to the CommonVoice website code included not only its localization into Welsh but also making the website functionality consistent with smaller prompt sets (i.e. to avoid providing already recorded prompts), additional profile meta-data fields, and mandatory provision of such profile meta-data before allowing recording.

To date there has not been a large publicity campaign to recruit further volunteers to record their voices through any Paldaruo interface as these are costly and time-consuming to organise. Just keeping the Paldaruo app active from the end of the original 2014 to late 2017 gathered around 136 additional recordings. However, a new appeal disseminated through social media sites and e-mails has drawn a good initial reaction, with further dissemination undertaken by volunteers retweeting, and incorporating the appeal in their own newsletters and e-mail lists. Participants who are reluctant to record their own voices now also feel able to contribute by verifying the accuracy of recordings made by others. We believe that an active publicity campaign would enable us to significantly increase both the number and length of contributions so far made.

4. Making results visible and accessible

Good visibility of the outputs made possible by crowdsourcing generates an enthusiasm amongst the public interested in supporting the development of Welsh language speech technology. All projects contain outreach activities which have been instrumental in generating a following, with past contributors keen to view and listen to the latest developments.

On a technical level, and as stipulated in the grants awarded, all project outputs to date have been published and shared with permissive open source licenses on the Welsh National Language Technologies Portal (Prys and Jones, 2016b), and/or on widely used repositories such as GitHub. This has led the team to engage with several developers, companies and enthusiasts interested in utilising the speech technology outputs in their own Welsh language provisions. This includes electronics students attempting to create a 'body' for Macsen as a personal assistant, and teacher trainers developing primary school lessons on coding in the context of Welsh language (Prys and Jones, 2017).

Further attempts at reaching out to developers and users have led to a website dedicated to supporting anyone wanting to obtain, create and develop their own Macsen. Based on a translation and fork of the Jasper project, the website can be found at https://projectmacsen.github.io as seen in Figure 2.

The widest audiences have been reached however via television and radio interviews in the Welsh language media when questions arose with the appearance and increasing popularity of Amazon Alexa and Google Home as to why there were no Welsh language versions. In the meantime, team members continue to present the work at societies and cultural events such as Hacio'r Iaith (*'Hacking the Language'*) the National Eisteddfod. Each outreach occasion presents opportunities to appeal for contributions to the Paldaruo speech corpus.

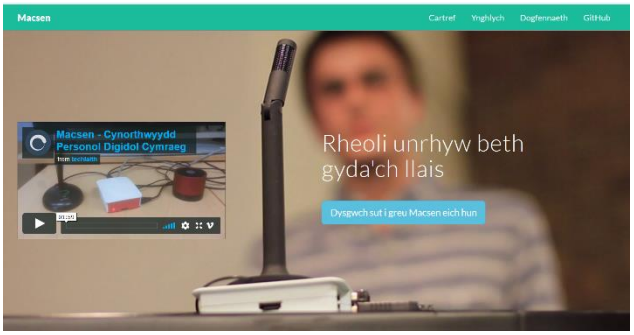

Figure 2 - Screenshot of website for Macsen users and developers (https://projectmacsen.github.io)

5. Conclusions and Future Work

No research project is undertaken in isolation. At their best, projects are undertaken to increase the sum of human knowledge and support human life and culture. Crowdsourcing brings researchers and their public closer together, and the greater intimacy offered by small language communities can bring positive benefits, not least in the field of speech technology.

On the other hand, generic, global projects, such as Wikipedia, MaryTTS and Mozilla Common Voice, offered on generous open licences, are of vital importance to less-resourced languages, enabling knowledge and resources to be shared and built up to the benefit of all. Combining use of local, language-specific knowledge and resources with global tools and initiatives can provide the help needed to level the playing field for digitally excluded language communities, and such combinations are to be welcomed.

As with speech technology for many other less-resourced languages, much remains to be done for Welsh. Within the current project, as well as gathering and processing additional data, improved acoustic and language models need to be built. Assessing quality, and answering the question 'how much data is enough, or at least sufficient' is becoming increasingly urgent as we seek to improve on the first on the first generation of outputs.

The increasing pace of technological developments, especially neural networks for speech recognition, is creating new challenges for less-resources languages, especially as truly enormous datasets are needed to gain the best results. However, smarter ways of working, use of both generic global and local language-specific knowledge can provide a way ahead for many less-resourced languages. We feel privileged to be part of such global and local communities.

6. Acknowledgements

The projects reported on in this paper were made possible with the financial support of the Welsh Government, through its Technology and Digital Media in the Welsh Language Fund and S4C. The authors would also like to thank the contributors from various hackers and communities of users that assisted us on the projects, as well as Robin Owain, Wikimedia UK Manager in Wales and Michael Henretty from the Mozilla Common Voice project for their aid.

7. References

Cooper, S. Jones, D. B. and Prys, D. 2014. *Developing further speech recognition resources for Welsh*. In: Judge, J., Lynn, T., Ward, M. and Ó Raghallaigh, B. eds. Proceeedings of the First Celtic Language Technology Workshop at the 25th International Conference on Computational Linguistics (COLING 2014), 23 August 2014, Dublin, Ireland. pp. 55-59.

DFKI. 2016. http://mary.dfki.de/ [accessed 12/01/2018]

Gorman, Kyle, Jonathan Howell and Michael Wagner. 2011. Prosodylab-Aligner: A Tool for Forced Alignment of Laboratory Speech. Canadian Acoustics. 39.3. 192–193.

Jones, D.B. and Cooper, S. 2016 *Building Intelligent Digitial Assistants for Speakers of a Lesser-Resourced Language*. p74-79 Proceedings of the LREC 2016 Workshop "CCURL 2016 – Towards an Alliance for Digital Language Diversity", Claudia Soria, Laurette Pretorius, Thierry Declerck, Joseph Mariani, Kevin Scannell, Eveline Wandl-Vogt.

Mozilla (n.d.) Common Voice https://voice.mozilla.org/. [accessed 12/01/2018]

Prys, D., Prys G., and Jones, D.B. 2016a. *Cysill Ar-lein: A Corpus of Written Contemporary Welsh Compiled fro m an On-line Spelling and Grammar Checker*. Proceedings of the Tenth International Conference on Language Resour ces and Evaluation (LREC 2016) Portoroz, Slovenia.

Prys, D., and Jones D. B. 2016b. *National Language Technology Portals for LRLs: A Case Study.* Language Technologies in Support of Less-Resourced Languages, (LRL 2015).

Prys, D., Jones, D.B & S. Ghazzali. 2017. *Using LT tools in classroom and coding club activities to help LRLs* Language Technologies in Support of Less-Resourced Languages, (LRL 2017).

The Mozilla Blog (2017) https://blog.mozilla.org/blog/2017/11/29/announcing-the-initial-release-of-mozillas-open-source-speech-recognition-model-and-voice-dataset/ [accessed 12/01/2018]

Voxforge http://www.voxforge.org/ [accessed 06/03/2018].

Welsh Government. 2013. Welsh language Technology and Digital Media Action Plan. http://gov.wales/docs/dcells/publications/230513-action-plan-en.pdf [accessed 12/01/2018]

Welsh Government. 2017. Cymraeg 2050: A million Welsh speakers. Work programme 2017-2021. http://gov.wales/docs/dcells/publications/170711-cymraeg-2050-work-programme-eng-v2.pdf [accessed 12/01/2018]