Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology

Cooper, Sarah; Jones, Dewi Bryn; Prys, Delyth

Information

DOI:
10.3390/info10080247

Published: 25/07/2019

Publisher's PDF, also known as Version of record

information

MDPI

Article

Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology

Sarah Cooper [1,*], Dewi Bryn Jones [2] and Delyth Prys [2]

[1] School of Languages, Literatures and Linguistics, Bangor University, Bangor, Gwynedd LL57 2DG, UK
[2] Language Technologies Unit, Bangor University, Bangor, Gwynedd LL57 2DG, UK
* Correspondence: s.cooper@bangor.ac.uk

Received: 30 June 2019; Accepted: 23 July 2019; Published: 25 July 2019

Abstract: Collecting speech data for a low-resource language is challenging when funding and resources are limited. This paper describes the process of designing, creating and using the Paldaruo Speech Corpus for developing speech technology for Welsh. Specifically, this paper focuses on the crowdsourcing of data using an app on smartphones and mobile devices, allowing speakers from across Wales to contribute. We discuss the development of reading prompts: isolated words and full sentences, as well as the metadata collected from contributors. We also provide background on the design of the Paldaruo App as well as the main uses for the corpus and its availability and licensing. The corpus was designed for the development of speech recognition for Welsh and has been used to create a number of other resources. These methods can be extended to other languages, and suggestions for other low-resource languages are discussed.

Keywords: low-resource languages; linguistic diversity; speech recognition; speech technology; corpus

1. Introduction

This paper describes the design and collection of the Paldaruo Speech Corpus of Welsh for developing speech technology. The corpus is crowdsourced from volunteers across Wales using the Paldaruo smartphone app ("paldaruo" means "chattering" in Welsh).

Welsh is a Celtic language primarily spoken in Wales in the United Kingdom and has been in long-term contact with English for hundreds of years. There are over half a million speakers of Welsh in Wales equating to around 19% of the population [1]. Welsh has a fairly strong presence in the media with Welsh radio stations and a TV channel (S4C) that broadcasts in Welsh [2]. Welsh is a minority language in Wales, and speakers are also bilingual in English. With English dominating the technological infrastructure in the UK, advances in language technologies in Welsh have been side-lined. It is challenging to develop computational resources for a language without established training data, especially given that there is no immediate or direct commercial benefit for private sector companies. During this project, the importance of the support of the language community themselves, as well as public organizations or foundations has become apparent in order to support language technology development in a low-resource language.

In this paper, the remainder of Section 1 provides the background for the paper. We present some of the challenges for low-resource languages in developing computational resources, provide background on variation in spoken Welsh and discuss previous work on Welsh speech technology. Section 2 describes the considerations for data collection and the design of the corpus including details of the Paldaruo app used to crowdsource the data from volunteers across Wales. Section 3 provides an overview of the corpus, the technical specifications of the data and presents uses of the corpus so far.

In Section 4 we discuss the challenges faced in the process of data collection and the implications of the design and collection method for other low-resource languages.

1.1. Challenges for Speech Technology and Low-Resource Languages

Natural language processing and artificial intelligence for building language technologies (e.g., translation or speech technology) play a role in the lives of billions of people worldwide. On the one hand, they may be seen as a threat to the survival of small languages such as Welsh, and on the other hand they can provide an opportunity for them to prosper in the contemporary digital world.

Several high-resource languages such as English, French, Spanish, Mandarin and Japanese dominate the commercial market [3,4]. As a result, small and low-resource languages may find themselves marginalized in terms of the availability of digital resources, leaving speakers at risk of being excluded economically and socially. The importance of electronic technologies in smaller languages is argued to be essential for ensuring a language progresses and avoids digital extinction [5–7]. The risk of digital extinction to smaller languages has featured extensively in the European Digital Language Diversity Project (DLDP). One output of the project is the Digital Language Survival Kit [8], which provides guidance and outlines some of the basic language technologies that small languages require to ensure their digital vitality.

Languages with small numbers of speakers often find themselves low-resourced with regards to the availability and interest in funding [8]. Speech recognition systems require a large number of speakers as well as a large data set in order to train accurate acoustic models for large vocabulary speech recognition. However, paying speakers to contribute their voices to speech corpora can often be costly in terms of compensation and the time required by researchers to perform the data collection. Furthermore, variation is desired which can often be an issue when speakers of a minority language are spread over a large geographical area. In order to collect data from a wide range of speakers from the major Welsh dialect areas, a crowdsourcing approach was adopted for the collection of data in the current project. Crowdsourcing involves nonexpert volunteers participating in an online activity, usually without compensation [9,10]. Crowdsourcing is viewed as a solution to the large data dilemma by speech and language processing communities and has the potential to vastly simplify audio collection efforts [10,11].

1.2. Spoken Welsh

Linguistic research into Welsh is fairly well established, with a tradition of research considering language variation and change in Welsh. There are 29 consonants and up to 13 monophthongs and 13 diphthongs dependent on the variety [12–15]. Welsh has traditionally been categorized as having six main dialect areas, shown in Figure 1 below. Welsh orthography is fairly transparent, with 29 graphemes (including six digraphs) used to represent the consonants and vowels.

Processes of dialect levelling as well as sociolinguistic variation at the level of phonetics and phonology, as well as with morphology and syntax, have been identified in recent years [16–22]. The consonant inventory across different dialect areas is fairly similar, however, the vowel inventories of speakers broadly from North and South Wales are fairly different. Due to the substantial variation present in speakers from different dialect areas in Wales, in planning of the Paldaruo Corpus, it was considered essential to collect data from speakers from the main dialect areas in order to develop resources for speech technology that would be appropriate for speakers from across Wales.

Figure 1. Main dialect areas of Wales (adapted from Thomas and Thomas [16] (p. 28)).

1.3. Welsh Language Technologies

The presence of Welsh in the digital environment is a critical factor in the widespread and continued use of the language outlined in the Cymraeg 2050: A million Welsh speakers strategy [23]. Following from the strategy, in 2018 the Welsh Government also released the Welsh language technology action plan [24] following on from an earlier plan [25]. In this most recent action plan, the Welsh Government outlined three specific areas to be addressed by the most recent plan:

1. Welsh language speech technology.
2. Computer-assisted translation.
3. Conversational artificial intelligence.

The plan sets out challenges for these areas and outlines principles to ensure sustainable provision for future generations.

Substantial work on Welsh speech technology was developed under the WISPR (Welsh and Irish Speech Processing Resources) project [26]. Previous work on a diphone-based synthesizer [27,28] and also a small speech database for Welsh [29] was built upon by the WISPR project. An improved synthetic Welsh voice was developed as part of the WISPR project as well as a MSAPI interface to Festival for use in Microsoft Windows environments. MSAPI is a Speech Application Programming Interface in an API specified by Microsoft (hence MSAPI) that can be implemented by speech recognition and speech synthesis components to facilitate their use by Windows applications. Further work into developing commercial Welsh synthetic voices was undertaken by the Language Technologies Unit at Bangor University, by the Finnish company Bitlips and the Polish company Ivona. The Basic Welsh speech recognition project at the Language Technologies Unit at Bangor University in 2008–2009 resulted in laboratory prototypes for (a) a "command and control" application for a PC where applications could be launched by voice control and (b) a simple voice-driven calculator.

The Paldaruo Speech Corpus was developed as part of the GALLU (Gwaith Adnabod Lleferydd Uwch-Further Speech Recognition Work) project funded by the Welsh Government and S4C, the

Welsh language publisher-broadcaster. The year-long project aimed to build on the Basic Welsh Speech Recognition project described above by developing resources for the development of speech recognition. Subsequent projects have aimed to develop speech recognition as components of a Welsh intelligent digital assistant which are described in Section 2.2 below.

So far, we have presented an introduction to the central considerations when designing and collecting a speech corpus for a low-resource language for speech technology. In order to reach the aim of developing speech recognition for Welsh, we specified the following main goals for the Paldaruo Corpus:

- To collect data from speakers from all major Welsh dialect areas, including data from native and non-native speakers of Welsh.
- To collect data which covers a representative sample of the most common sounds in the language.

The remainder of this paper will outline the data design and collection, provide an analysis of the structure of the corpus and the main uses of the corpus thus far, followed by a discussion of the experience of data collection, current and future data collection aims and recommendations for crowdsourcing speech data in other languages.

2. Materials and Methods

2.1. The Paldaruo App

In order to crowdsource data from speakers across Wales, the Paldaruo App [30] was designed for iOS and Android devices. Rather than relying on traditional methods of paying speakers to record in a sound booth, mobile devices with inbuilt microphones and internet connectivity are ideal for crowdsourcing data for a large speech corpus. The app was designed to be simple to use in order to maximize the number of contributions from each volunteer. Each volunteer creates a profile whereby they fill in their language background, including information about their age, gender, the area in which they spent their childhood, the area in which they currently live, how often they speak Welsh and with whom they speak Welsh. This data is attached to a randomly generated user ID. The app also contains a video, accessible from the home screen, which provides the user with an introduction and a demonstration of the recording process.

Due to the fact that speech data was being collected, ethical approval was sought from Bangor University's College of Arts and Humanities Ethics Committee. After creating their profile, volunteers are provided with details of the intended uses of the data, and consent to their contributions being collected and used for speech research. In order to comply with research ethics guidelines, we also do not publish data from volunteers under the age of 18.

After creating their profile, volunteers can begin recording the prompts (see Section 2.2 below). The app accesses the microphone of the user's mobile device and records 48 kHz PCM files. The volunteer records each prompt individually, and the recording is replayed to the volunteer who verifies the quality or re-records. Once approved, the recording is sent to a secure server hosted by the Language Technologies Unit at Bangor University. The uploads are queued in the background so that network speed issues do not interrupt the recording process.

The app was designed in order to allow the volunteers to stop and resume recording at any time. In order to ensure coverage of as many of the items as possible, the prompts were provided to each volunteer in a different random order.

Studies using crowdsourcing methodology must pay attention to quality control in the data. We implemented methods to ensure that the contributed data was of adequate quality for the development of speech recognition. During a pilot use of the app with a small number of users, it was evident that some users were unsure of how far to hold the mobile device away from their mouths resulting in recordings that were either too quiet, or too loud and therefore clipped. It also was evident during pilot use that the levels of background noise varied between office and home settings. Therefore,

we introduced a sound level check that measured the audio's volume levels at various stages in the recording process. After the check, the volunteer would be informed if the recording was too quiet or too loud. This feature was implemented by the app taking measurements during the recording, at intervals of 0.03 s, of the audio signal's peak power in decibels (a function provided by the device's underlying operating system e.g., iOS AVFoundation–AVAudioRecorder–peakPowerForChannel) and updating a register of the highest so far. This formula converts the measurements in decibels to an amplification gain or volume level:

$$amplification\ gain\ =\ 10^{\frac{\max(db)}{20}} \quad (1)$$

Thus, by measuring maximum volume level and checking against a desired amplification gain between 0.5 and 1.0, the app is able to determine if a recording is too quiet or too loud and inform the user of whether to speak louder or hold the device further away. In addition, the app insists on measuring background noise before beginning each recording session by asking for a recording without the user speaking. If the result measurement has a maximum volume level higher than 0.5 then the app informs the user that the background noise is too high and asks the user to change their environment. We aimed to ensure that the app collected data that would be of appropriate quality for the development of speech recognition using these quality control methods.

2.2. Prompt Design

Given the challenges for low-resource languages described above, the prompts to be read out for the first phase of the corpus were designed to capture the most frequent sound combinations in the language. When designing the prompts, we checked that the words were easily readable to ensure volunteers would be familiar with the items. The first phase of the corpus contained 43 prompts consisting of 8 individual words per prompt. Examples of the individual word prompts are shown in Table 1 below, with English translations. In total there were 344 words in the first phase of the corpus, which took contributors around 30 min to record.

Table 1. Examples of individual word prompts with English translations.

Prompt	Translation
lleuad, melyn, aelodau, siarad, ffordd, ymlaen, cefnogaeth, Helen	moon, yellow, members, talk, road, forward, support, Helen
gwraig, oren, diwrnod, gwaith, mewn, eisteddfod, disgownt, iddo	wife, orange, day, work, in, eisteddfod, discount, to him
rhybuddio, Elen, uwchraddio, hwnnw, beic, Cymru, rhoi, aelod	warn, Elen, upgrade, that, bike, Wales, give, member
lliw, yng Nghymru, gwneud, rownd, ychydig, wy, yn, llaes	colour, in Wales, make, round, few, egg, in, flaccid
hyn, newyddion, ar, roedd, pan, llun, melin, sychu	this, news, on, was, when, picture, mill, dry

Additional prompts were added to the corpus in 2015 as part of a subsequent project to develop speech and language resources for Welsh and in order to develop a prototype Welsh digital assistant "Macsen" [31]. The additions aimed to expand the data to include full sentences rather than individual words in order to ensure that a greater variety of data was collected. Examples of the question and sentence prompts are shown in Table 2 below, with English translations. Again, in order to ensure readability, these were carefully checked and developed based on a prompt set from the MaryTTS resources developed in an earlier project. We added sentences that would facilitate answering questions to a digital personal assistant such as "What is the time?", "What is the weather for today?", "What's the news?", "Play me some music", "Play me some Welsh music". In order to expand the capabilities further, we also expanded the data set to include questions from some of the top searched Wikipedia articles, as well as adding more general reading prompts (for further details, see [31,32]).

Table 2. Examples of full question/sentence prompts with English translations.

Prompt	Translation
Beth ydy Cymraeg?	What is Welsh?
Beth oedd Yr Ail Ryfel Byd?	What was the Second World War?
Pwy oedd T. Llew Jones?	Who was T. Llew Jones?
Faint mae llaeth yn costio?	How much does milk cost?
Daeth wyau siocled yn boblogaidd adeg oes Fictoria	Easter eggs became popular during the Victorian period
Doedd dim cerbyn arall yn rhan o'r ddamwain	No other vehicles were involved in the accident
Mi fydd y broses yn debyg i etholiadau eraill	The process will be similar to other elections
Dw i wrth fy modd yn cerdded ac yn hoff iawn o natur	I'm in my element walking and am fond of nature
Mae unrhyw ddraenog sydd allan yng ngolau dydd angen help	Any hedgehog that is out in daylight needs help

2.3. Data Collection

The Paldaruo app was launched on iOS and Android on 7 July 2014. We raised awareness of the app via social media and press releases. In order to raise the profile of the story, Carwyn Jones, Wales' First Minister at the time, recorded his voice and his contribution was advertised as the first to respond to the appeal for volunteers. These efforts resulted in several news articles [33–35] and reporting on the radio and television news programs. The authors also appeared on the BBC Radio Cymru Post Cyntaf program on the morning of the launch, and on the S4C television program Heno in the evening. The app was downloaded 181 times on the day of the launch and since then has been downloaded 628 times on iOS and 207 times on Android, resulting in a total number of downloads of 835 since the launch.

3. Corpus Analysis

This section describes the breakdown of the versions of the Paldaruo Corpus as well as the uses of the data for developing speech resources. The first version of the corpus was published in November 2014, 5 months after the launch of the Paldaruo App. As shown in Table 3 below, the majority of contributions were made in these first few months. In 2016, an updated Version 2 of the corpus was made available, where a further 8 h of data were added. Since 2016, regular versions of the corpus have been published showing a slow but steady increase in the amount of data from an increasing number of volunteers.

Table 3. Published versions of the Paldaruo Speech Corpus.

	Version 1	Version 2	Version 3	Version 4	Version 5
Date Published	31 November 2014	15 July 2016	9 June 2017	16 November 2017	19 December 2018
Audio Duration (hours)	26	34	36	38	40
Number of files	8941	11,556	12,024	12,682	14,215
Number of contributors	383	487	506	536	564

In terms of the demographic information about the volunteers of the corpus, Figure 2 below shows the age and gender distribution in the data. We published contributions from a range of speaker age groups, with the youngest speaker being 18 and the oldest speaker being 80. The majority of volunteers were in the 20–29 age group (n = 168) and the 30–39 age group (n = 171). There is a fairly even split in terms of the gender distribution in the data: overall in Version 5 of the corpus, 49.3% of the volunteers were male and 50.5% of volunteers were female.

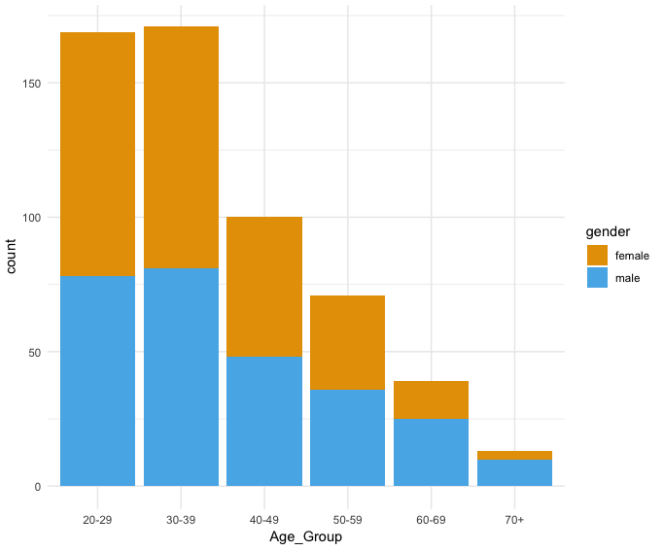

Figure 2. Age group and gender distribution of volunteers in Version 5 of the Paldaruo Corpus.

Figure 3 below shows the distribution of volunteers from the different dialect areas of Wales, with information about whether volunteers categorized themselves as having a first language accent, or a learner accent. In terms of the distribution of speakers from different areas, the majority of volunteers were from north west Wales (n = 239, 42.4% of the data), which corresponds with the highest concentration of speakers of Welsh according to the most recent census data [1]. Volunteers from south west Wales made up around 19% of the data (n = 108), followed by volunteers from south east Wales (n = 82, 14.5%), mid-Wales/other (n = 81, 14.4%), with the smallest number of volunteers coming from north east Wales (n = 53, 9.4%). Volunteers were asked to categorize their accent as first language or learner accent. Eighty-six percent of the speakers considered themselves to have a first language accent, while the remaining 14.0% considered themselves to have a learner accent.

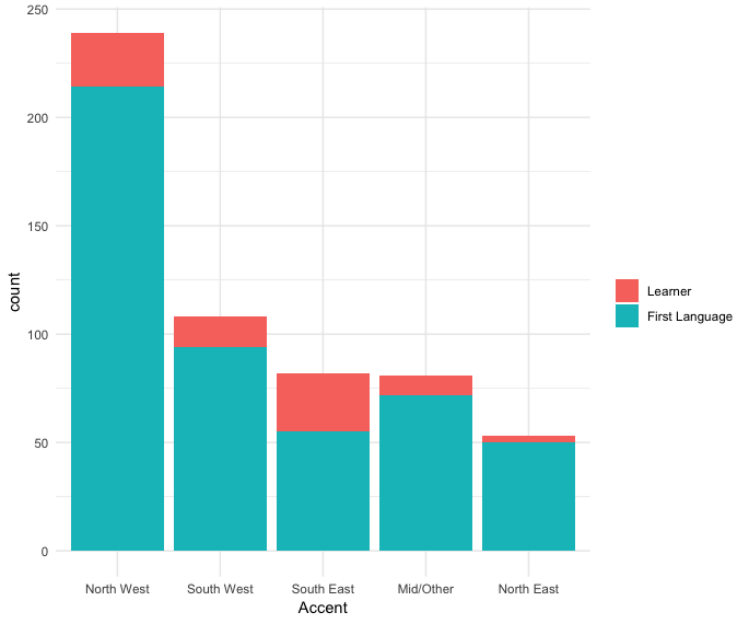

Figure 3. Accent distribution of volunteers in Version 5 of the Paldaruo Corpus.

Technical Specifications and Use of the Paldaruo Corpus

The corpus is distributed under the CC BY license for academic and commercial use via META-SHARE and the Welsh National Language Technologies Portal [36]. All acoustic data in

the corpus are encoded in wav format at 48 kHz, 16 bit Linear PCM, mono. The labels of the spoken prompts in each sample are stored in a single plain text file. The anonymized metadata of participants are available in a comma-separated values file which lists the unique id number, and the other metadata described in Section 2.2.

The Welsh National Language Technologies Portal also houses a wide range of resources for the Welsh language including resources for Welsh language translation, speech resources, corpora, online APIs service and website plugins. As stipulated by the grants awarded for the development of these resources by the Welsh Government and S4C, outputs are published and shared with open source licenses.

Since 2014, work carried out using the Paldaruo corpus has focused on developing speech recognition and associated resources. Doctoral work [37,38] beginning in 2016 has been focusing on developing speech recognition for Welsh using different toolkits including HTK, Kaldi and Mozilla's DeepSpeech [39–41]. The work also focusses on differences in the accuracy of the systems in responding to test sets from different dialect areas in Wales. The results of this project will be available in early 2020. As part of the GALLU project, the speech data was successfully used to build acoustic models that recognized commands to move a robot arm connected to a Raspberry Pi using spoken commands in Welsh [42]. As part of subsequent projects, packages for speech recognition have been built in Docker-based environments for users to easily produce, test and apply acoustic models in Kaldi [40], HTK [39] and Julius [43] using the Paldaruo Corpus. Subsequent projects funded by the Welsh Government aimed to develop a Welsh digital intelligent assistant "Macsen", which incorporates speech recognition, and text-to-speech [31,32].

Acoustic models developed using the Paldaruo Corpus were used with the Prosodylab-Aligner [44]. The forced aligner automatically aligns and specifies every word and phoneme within a sound file. Traditionally, alignment for phonetic research is tackled by manually matching the text with the sound recordings. However, as some speech corpora are very large, automatic methods are needed to align text with speech for large scale data analysis. Figure 4 below shows the output from Praat [45] indicating the forced alignment of the individual sounds and words in an audio recording from the spoken first line of the Welsh national anthem.

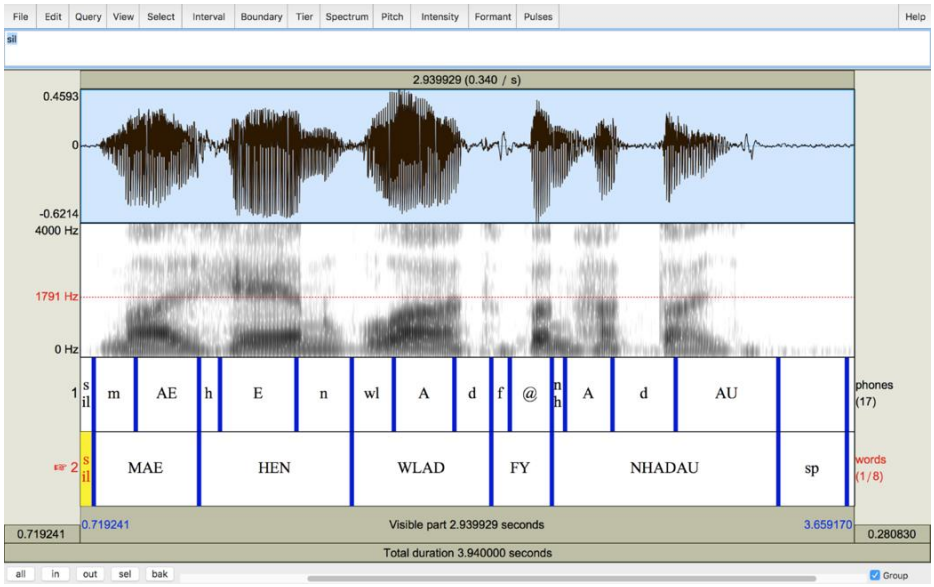

Figure 4. Praat output showing forced alignment for Mae hen wlad fy nhadau, the first line of the Welsh national anthem Land of my fathers.

The figure indicates that the aligner has found the word boundaries and has fairly accurately aligned the individual phones within the words. This process reduces the amount of human annotation

of the time-boundaries for phonetic data analysis. The forced alignment can be checked by a human labeller for accuracy and adjusted as required, speeding up the process of data analysis.

The Paldaruo Corpus has primarily been used for the development of speech recognition for Welsh, although it has also been used as a resource for researchers interested in the phonetics and phonology of the language, with further work ongoing [46,47].

4. Discussion

Language-specific speech technologies such as speech recognition are dependent on large speech corpora consisting of speakers' spoken examples with associated transcriptions. The motivation of this research work came from the need to ensure that a large labelled speech corpus was available for the development of speech technologies for Welsh, which is supported by the Welsh Government and in their strategy to reach 1 million speakers of Welsh by 2050 [23].

We have presented the process of designing and collecting the Paldaruo Speech Corpus, as well as providing details about the distribution of speakers in the corpus. We aimed to collect data which covered a representative sample of the most common sounds in the language, which was achieved through designing recording prompts for volunteers to read out loud. We also collected data from speakers from all of the major Welsh dialect areas, including data from native and non-native speakers of Welsh. This was achieved through promotion and publicizing of the release of the Paldaruo app on social media and in press outlets. The occasional promotion of the campaign to develop speech recognition over the past five years has resulted in further contributions resulting in a small but steady increase in the number of contributions. The most recent version of the corpus contains 40 h of data from 564 speakers, and this data has been used in ongoing projects to develop speech recognition for Welsh.

In terms of other languages, this case study demonstrates that crowdsourcing methods using mobile devices can be effective for developing speech and language resources for computational methods in low-resource languages [11]. The collection of the Paldaruo Corpus capitalized on the public interest in developing speech technologies for Welsh. During this project, we found that effective publicizing of efforts was essential in ensuring that the language community was aware of the benefits of their contribution. The source code for the iOS app [48] and accompanying server component is available for other developers or language communities.

In terms of developments using the data in the future, the Paldaruo app continues to collect contributions for the development of speech technology. Since 2017, the authors have been working with Mozilla with its CommonVoice initiative at crowdsourcing speech data in a similar way. Initially, CommonVoice was launched with English only. Through working with Mozilla, the experiences of collecting the Paldaruo Corpus for Welsh illustrated that collecting speech data from low-resource speech communities was viable and necessary to meet their vision to ensure that the Internet is a global public resource, open and accessible to all. In 2018 Mozilla expanded the number of languages offered. A multilingual version of CommonVoice was launched in June 2018 to crowdsouce data for German, French and Welsh. For Welsh, the prompts from the Paldaruo app were included on the platform. The CommonVoice initiative has since been expanded to 13 other languages, including other low-resource languages such as Breton and Catalan, with 76 other languages under development. As a result of the data collection methods, in January 2019, Mozilla published 22 h of Welsh speech data, followed by a second release of the corpus with 47 h in June 2019.

In conclusion, we have illustrated that crowdsourcing speech data using smartphones is a useful method for low-resource languages to develop resources for computational methods [10,11]. Given the support of a few high-resource languages in commercial companies, smaller languages who wish to ensure digital vitality may source data from the language community to collect low-cost large data sets for the development of language technologies [3,5,6,8].

Author Contributions: Conceptualization, D.B.J. and D.P.; Data curation, D.B.J.; Formal analysis, S.C. and D.B.J.; Funding acquisition, D.P.; Methodology, S.C. and D.B.J.; Project administration, S.C. and D.P.; Resources, D.B.J.;

Software, D.B.J.; Validation, D.B.J.; Visualization, S.C.; Writing—original draft, S.C.; Writing—review & editing, S.C. and D.B.J.

Funding: This research was funded by The Welsh Government under the Welsh-language Technology and Digital Media Grant and S4C.

Acknowledgments: We thank David Chan, research officer on the original GALLU project for assistance with data design, collection and software. We wish to thank the volunteers who have and continue to contribute their speech data.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the Paldaruo Corpus, the writing of the manuscript or in the decision to publish the results. S4C supported the data collection via promotion with news articles and on television programs.

References

1. *Language in England and Wales: 2011*; Office for National Statistics: South Wales, UK, 2013.
2. Aitchison, J.W.; Carter, H. *A Geography of the Welsh Language, 1961–1991*; University of Wales Press: Cardiff, UK, 1994; ISBN 978-0-7083-1236-0.
3. Besacier, L.; Barnard, E.; Karpov, A.; Schultz, T. Automatic speech recognition for under-resourced languages: A survey. *Speech Commun.* **2014**, *56*, 85–100. [CrossRef]
4. Kurimo, M.; Enarvi, S.; Tilk, O.; Varjokallio, M.; Mansikkaniemi, A. Modeling under-resourced languages for speech recognition. *Lang. Resour. Eval.* **2017**, *51*, 961–987. [CrossRef]
5. Crystal, D. *Language Death*; Cambridge University Press: Cambridge, UK, 2014.
6. Cormack, M. The media and language maintenance. In *Minority Language Media: Concepts, Critiques and Case Studies*; Cormack, M., Hourigan, N., Eds.; Multilingual Matters: Clevedon, UK, 2007; pp. 52–68.
7. Pretorius, L.; Soria, C. Introduction to the special issue. *Lang. Resour. Eval.* **2017**, *51*, 891–895. [CrossRef]
8. Ceberio Berger, K.; Gurrutxaga Hernaiz, A.; Baroni, P.; Hicks, D.; Kruse, E.; Quochi, V.; Russo, I.; Salonen, T.; Sarhimaa, A.; Soria, C. *Digital Language Survival Kit: The DLDP Recommendations to Improve Digital Vitality*. Available online: http://wp.dldp.eu/wp-content/uploads/2018/09/Digital-Language-Survival-Kit.pdf (accessed on 24 July 2019).
9. Estellés-Arolas, E.; González-Ladrón-de-Guevara, F. Towards an integrated crowdsourcing definition. *J. Inf. Sci.* **2012**, *38*, 189–200. [CrossRef]
10. Eskenazi, M. The basics. In *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*; Eskenazi, M., Levow, G., Meng, H., Parent, G., Suendermann, D., Eds.; Wiley: Hoboken, NJ, USA, 2013.
11. McGraw, I. Collecting Speech from Crowds. In *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*; Eskenazi, M., Levow, G., Meng, H., Parent, G., Suendermann, D., Eds.; Wiley: Hoboken, NJ, USA, 2013.
12. Jones, G. The distinctive vowels and consonants of Welsh. In *Welsh Phonology: Selected Readings*; Ball, M.J., Jones, G., Eds.; University of Wales Press: Cardiff, UK, 1984; pp. 40–64.
13. Mayr, R.; Davies, H. A cross-dialectal acoustic study of the monophthongs and diphthongs of Welsh. *J. Int. Phon. Assoc.* **2011**, *41*, 1–25. [CrossRef]
14. Ball, M.J.; Williams, B.J. *Welsh Phonetics*; Edwin Mellen Press: New York, NY, USA, 2001.
15. Awbery, G.M. Phonotactic constraints in Welsh. In *Welsh Phonology: Selected Readings*; Ball, M.J., Jones, G., Eds.; University of Wales Press: Cardiff, UK, 1984; pp. 65–104.
16. Rees, I.W. Phonological Variation in Mid-Wales. *Stud. Celt.* **2015**, *45*, 149–174.
17. Mayr, R.; Morris, J.; Mennen, I.; Williams, D. Disentangling the effects of long-term language contact and individual bilingualism: The case of monophthongs in Welsh and English. *Int. J. Biling.* **2017**, *21*, 245–267. [CrossRef]
18. Durham, M.; Morris, J. (Eds.) *Sociolinguistics in Wales*; Palgrave Macmillan: London, UK, 2017; ISBN 978-1-137-52897-1.
19. Morris, J. Sociophonetic variation in a long-term language contact situation: /l/-darkening in Welsh-English bilingual speech. *J. Socioling.* **2017**, *21*, 183–207. [CrossRef]
20. Prys, M. Style in the vernacular and on the radio: code-switching and mutations as stylistic and social markers in Welsh. Ph.D. Thesis, Prifysgol Bangor University, Bangor, UK, 2016.

21. Davies, P.; Deuchar, M. Auxiliary deletion in the informal speech of Welsh–English bilinguals: A change in progress. *Lingua* **2014**, *143*, 224–241. [CrossRef]
22. Borsley, R.D.; Tallerman, M.; Willis, D. *The Syntax of Welsh*; Cambridge Syntax Guides; Cambridge University Press: Cambridge, UK, 2007.
23. Welsh Government. *Cymraeg 2050: A Million Welsh Speakers*; Welsh Government: Cardiff, UK, 2017.
24. Welsh Government. *Welsh Language Technology Action Plan*; Welsh Government: Cardiff, UK, 2018.
25. Welsh Government. *Welsh-Language Technology and Digital Media Action Plan*; Welsh Government: Cardiff, UK, 2013.
26. Prys, D.; Williams, B.; Hicks, B.; Jones, D.B.; Ní Chasaide, A.; Gobl, C.; Carson-Berndsen, J.; Cummins, F.; Ní Chiosáin, M.; McKenna, J.; et al. WISPR: Speech Processing Resources for Welsh and Irish. In Proceedings of the SALTMIL Workshop at LREC 2004: First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, Lisbon, Portugal, 24 May 2004; pp. 68–71.
27. Williams, B. Diphone synthesis for the Welsh language. In Proceedings of the 1994 International Conference on Spoken Language Processing, Yokahama, Japan, 18–22 September 1994; pp. 739–742.
28. Williams, B. Text-to-speech synthesis for Welsh and Welsh English. In Proceedings of the Eurospeech 1995, Madrid, Spain, 18–21 September 1995; Volume 2, pp. 1113–1116.
29. Williams, B. A Welsh speech database: Preliminary results. In Proceedings of the Eurospeech 1999, Budapest, Hungary, 5–9 September 1999; Volume 5, pp. 2283–2286.
30. Language Technologies Unit. *Paldaruo*; Bangor University: Bangor, UK, 2019.
31. Jones, D.B.; Cooper, S. Building Intelligent Digital Assistants for speakers of a Lesser-Resourced Language. In Proceedings of the LREC 2016 Workshop "CCURL 2016—Towards an Alliance for Digital Language Diversity", Portorož, Slovenia, 23–28 May 2016; pp. 74–79.
32. Prys, D.; Jones, D.B. Gathering Data for Speech Technology in the Welsh Language: A Case Study. In Proceedings of the LREC 2018 Workshop "CCURL 2018—Sustaining Knowledge Diversity in the Digital Age", Miyazaki, Japan, 12 May 2018; pp. 56–61.
33. BBC. *Lansio adnodd Adnabod Lleferydd Cymraeg Newydd (Launching a New Welsh Speech Recognition Resource)*; BBC Cymru Fyw Website; BBC: London, UK, 2014.
34. BBC. *Speakers for Welsh Voice Recognition App Sought*; BBC News Website; BBC: London, UK, 2014.
35. S4C. *Apêl am Leisiau i Helpu Adeiladu Adnodd Adnabod Lleferydd Cymraeg (Appeal for Voices to Help Create Welsh Speech Recognition Resource)*; S4C News; S4C: Wales, UK, 2014.
36. Language Technologies Unit. *Welsh National Language Technologies Portal*; Bangor University: Bangor, UK, 2019.
37. Williams, I. *Challenges for Developing Speech Technology for Welsh*; Plas Gregynog: Newtown, UK, 2017.
38. Williams, I. *Modelau Cyfrifiadurol ar Gyfer y Gymraeg (Computational Models for Welsh)*; Bangor University: Bangor, UK, 2017.
39. Young, S.; Evermann, G.; Kershaw, D.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; Valtchev, V.; Woodland, P. *The HTK Book*; Version 3.2.; Cambridge University Engineering Department: Cambridge, UK, 2002.
40. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi Speech Recognition Toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Big Island, HI, USA, 11–15 December 2011.
41. *DeepSpeech: A TensorFlow Implementation of Baidu's DeepSpeech Architecture*; Mozilla: Mountain View, CA, USA, 2019.
42. Cooper, S.; Jones, D.B.; Prys, D. Developing further speech recognition resources for Welsh. In Proceedings of the First Celtic Language Technology Workshop at the 25th International Conference on Computational Linguistics, Dublin, Ireland, 23–29 August 2014; pp. 55–59.
43. Lee, A.; Kawahara, T. *Julius-Speech/Julius: Release 4.5*. Available online: https://zenodo.org/record/2530396#.XTgW4Y8RXIU (accessed on 24 July 2019).
44. Gorman, K.; Howell, J.; Wagner, M. Prosodylab-Aligner: A Tool for Forced Alignment of Laboratory Speech. *Can. Acoust.* **2011**, *39*, 192–193.
45. Boersma, P.; Weenink, D. *Praat: Doing Phonetics by Computer*. Available online: http://www.fon.hum.uva.nl/praat/ (accessed on 24 July 2019).

46. Cooper, S. *A Resource for Exploring Socio-Phonetic Variation in Welsh: The Paldaruo Corpus*; University of Glasgow: Glasgow, UK, 2015.
47. Iosad, P. *Bridging the Gap: Length and Tenseness in Brythonic Vowels*; Institiúid Ard-Léinn Bhaile Átha Cliath: Dublin, Ireland, 2017.
48. Language Technologies Unit. *Paldaruo Source Code*; Bangor University: Bangor, UK, 2019; Available online: https://github.com/techiaith/Paldaruo (accessed on 24 July 2019).