Proceedings of the Language Technologies for All (LT4All) , pages 367–370
Paris, UNESCO Headquarters, 5-6 December, 2019.

Planning for Language Technology Development and Language Revitalization in Wales

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

Abstract

Welsh speakers have striven to maintain and revitalize their language in the face of the increasing dominance of English. Language technologies, the internet and digital media important to make Welsh more attractive, relevant and accessible. From the initial efforts of a few academics, key resources and tools were created and formed into coherent building blocks to be reused and refined, keeping costs low and working with different stakeholders, including government, industry and enthusiasts. Recent publication of a government LT Action Plan has enabled longer term planning in ways which might also interest other language communities.

Keywords: language technology planning, best practice, Welsh

Résumé

Mae siaradwyr Cymraeg wedi ceisio cynnal ac adfer eu hiaith yn wyneb bygythiad cynyddol y Saesneg. Mae technolegau iaith, y rhyngrwyd a chyfryngau digidol yn bwysig i wneud y Gymraeg yn fwy deniadol, perthnasol a hygyrch. O ymdrechion cychwynnol rhai academyddion, crëwyd adnoddau ac offer allweddol a'u llunio yn flociau adeiladu i'w hailddefnyddio a'u gwella, gan gadw costau yn isel a gweithio gyda gwahanol randdeiliaid, gan gynnwys llywodraeth, diwydiant a charedigion yr iaith. Mae cyhoeddi Cynllun Gweithredu TI y llywodraeth yn ddiweddar wedi galluogi cynllunio tymor hirach mewn ffyrdd a allai fod hefyd o ddiddordeb i gymunedau ieithoedd eraill.

1. Introduction

Under-resourced languages are, by definition, lacking in adequate resources to fulfil their technological needs. It is common in such situations to start working on whatever the most urgent problem happens to be, such as a word list, bilingual dictionary, spell-checker or some other tool or resource. Before long, a handful of different tools and resources might have been created. They can fulfil a short-term need, but in order to be useful in the longer term, it helps that issues such as appropriate licences, sustainable repositories and modular design are all thought through as early as possible.

In the case of Welsh, there was an awareness of some of these issues from an early date, and attempts were made to future-proof the work in low-cost and sustainable ways.

Welsh is spoken by approximately 562,000 people (Office of National Statistics, 2012), and is variously described as minoritized or endangered, or more recently, in digital environments, a less- or under-resource language. After more than a century-long period of decline, there are concerted efforts, by Welsh-speakers, their non-Welsh speaking compatriots, and government, to reverse the decline and ensure the future of Welsh as a spoken, vibrant language, fit for life in the twenty first century.

2. Strategic documents

Since the devolution of power to the National Assembly for Wales in 1999, the government of Wales has published a succession of strategic documents to revitalize Welsh. These include Iaith Pawb: A National Action Plan for a Bilingual Wales (2003); A living language: a language for living (2012) and Cymraeg 2050: Welsh Language Strategy (2017). The latest of these has the ambitious target of nearly doubling the number of Welsh speakers to one million by the year 2050.

All these documents include sections on language technologies and their importance for the revitalization of Welsh. These sections have become increasingly focused and detailed as the technologies themselves have developed and assumed an increasingly central role in our lives over the last twenty years. Cymraeg 2050 emphasizes regional economic development and investing in entrepreneurship programmes to support Welsh speakers in rural areas. Digital technologies have a whole section devoted to them in this document and they also feature prominently in the section on linguistic infrastructure.

The Welsh language and technology have also been integrated into other strategic documents, policies and legislation, so that they do not exist in an isolated environment. Foremost amongst these has been the Well-being of Future Generations (Wales) Act (2015) with its vision to improve the social, economic, environmental and cultural well-being of the people of Wales, and the North Wales Growth Deal (2019) which names Technology as one of its priorities.

A Welsh Language Technology Action Plan (Welsh Government, 2018) added further detail these strategies, outlining "How we will ensure that more digital resources are available to support the use of Welsh" (Welsh Government, 2018).

This plan identified three specific areas to be addressed, namely:

1. Speech technology
2. Computer-assisted translation
3. Conversational Artificial Intelligence.

In each case the challenges are addressed, and in addition, underpinning themes are elaborated, including:

- Creating and sustaining digital infrastructure
- Developing a culture of open innovation
- Building capacity and digital skills

- Digital transformation in the public sector
- Promoting the creation and use of Welsh language digital products and services.

Welsh is therefore in a privileged position for a minoritized language in having a well-defined roadmap for future action.

3. Relevant Projects So Far

Several projects have been undertaken in recent years to lay the foundations for a coherent programme of language technology development for Welsh. Even without large scale, long term funding, academic researchers have been able to refer to government policies when making grant applications for the short-term projects then on offer. Previously there was no strong tradition of research in speech technology, machine translation or AI in Welsh universities, although it can be argued that these were in any case new fields of study, and that developing these areas for Welsh meant that Wales gained important new capacity in these fields. The main research so far has been conducted at Bangor University in north Wales, with some related activity in other institutions, and renewed efforts to establish an all-Wales research network in language technologies to advance the field in general.

Early projects concentrated on text-to-speech, since visually impaired Welsh speakers in the early 2000s were unable to access e-mails, text documents, and other materials on their computers that was written in Welsh. The ground-breaking WISPR (Welsh and Irish Speech Processing Resources) project was funded by the Interreg IIIA EU programme, leading to the first easy-to- use synthetic voices for Welsh, and later also for Irish (Williams, Prys and Ní Chasaide, 2005). Later research developed speech recognition resources for Welsh (Cooper, Jones and Prys, 2014; Prys and Jones, 2018 (1)), this time funded by the Welsh Government. Resources first developed in the WISPR project for text-to-speech, such as a Welsh pronunciation lexicon, were reused and updated, as part of a philosophy of making the best use of resources available.

In the meantime, advances in machine translation (MT) was bringing ever improving results for English and some other major languages. Minoritized languages often exist in bilingual environments alongside the dominant major language, and local translation industries have developed as a result. A report on translation tools for the translation industry in Wales (Prys, Prys and Jones, 2009) discussed MT tools for Welsh, leading to further research on MT for the Welsh-English language pair. Importantly, a Knowledge Transfer Partnership (KTP) project with a local translation company enabled that company to develop high quality domain specific MT using the company's own vast archive of legacy translations (Prys and Jones, 2019). Again, it was possible to share and reuse some resources such as wordlists with the speech technology projects.

The most recent addition to this mix of language technologies has been conversational artificial intelligence. As part of the 'Macsen' project to create a prototype personal assistant in Welsh (Jones and Cooper, 2016), spoken questions had to be understood and replied to appropriately. Although at a basic level this was possible by listening for some key words and using various APIs to provide answers relating to news, weather and time, to progress further in this field research is needed in intent parsing, natural language generation and many other new areas. This may seem overly ambitious for a small language, but in bilingual communities, where public authorities and private companies are moving towards AI conversational agents for reasons of cost and efficiency, the minoritized language has no option but to try and keep up. Again, reusing existing datasets and resources, refining them, and donating them back to the community go some way towards making such projects achievable.

The open-source platforms used include MaryTTS (Pammi et. al. 2010) for text to speech, Kaldi (Povey et. al. 2011) and later DeepSpeech (Mozilla) for speech recognition, and Moses-SMT (Koehn et. al. 2007) for MT. However, some of these platforms are challenging for others to use, and containerized wrappers have been developed at Bangor University for Moses and DeepSpeech (Jones, 2015 and 2018) to make this software more user friendly for non-experts. These platforms are useful for a large number of languages and contribute greatly to keeping down costs in developing LTs for less-resourced languages.

4. Licencing and Dissemination

The release of data itself under open source licences has been the subject of some debate in Wales as elsewhere. Large datasets are one of the core requirements to train any models in speech technology, MT and conversational AI agent applications. Finding enough appropriately licenced data is one of the biggest challenges for less-resourced communities. Any strategy for efficient and effective harvesting of data can make the difference between supporting a language or not in a software package. For example, attempts have been made with apps specifically developed to crowdsource a speech corpus from the Welsh language community (Cooper, Jones and Prys, 2019). More recent activity in crowdsourcing Welsh language speech data has focused on collaborating with and sharing efforts with Mozilla's CommonVoice initiative, since its philosophy and motives align (Prys and Jones (2018 (1)). Not all language communities are happy to lose control of their data, especially where they have had bad experiences of colonial exploitation in the past. However, in the Welsh context, the use of crowdsourcing strategies and of permissive licensing of data has helped the development of Welsh language software by the private sector, aided in some cases by knowledge transfer partnerships between academia and industry.

Even in Wales it has not always been possible to release data on open source licence, as some legacy products came with their own, previous licences. In other cases, there was the need to sell commercial software in order to fund the continuation of the work. Increasingly however, and wherever public funding was used to create the tools and resources, they were released on permissive licences such as BSD, MIT, Apache or CC-0, which permit reuse without any restrictions. This was in order to make the tools and

resources attractive to enable both small and large companies to take up and use in their own products. In both cases, the private sector is less willing to take up tools and resources published under copyleft licences, such as GPL and CC-BY-SA, which stipulate that the entire utilising body of software must be released openly under the same licence.

Although Welsh, with its approximately half a million speakers is deemed to be a very small market for commercial companies, it is still large enough to support many small companies who could benefit from language technology tools and resources. These include translation companies, local media, software companies, web designers, and producers of educational games and language teaching materials. In common with the experience of many other minoritized and endangered languages in peripheral regions, there are high proportions of Welsh speakers in the rural and remote north and west of Wales, areas that are impoverished with few opportunities for well-paid employment and therefore suffer from emigration of young, talented people. Providing appropriately licensed language resources to small, local companies in these areas can therefore help make them viable and help economic as well as linguistic revitalization of these areas.

The arguments for releasing resources on permissive licences for large multinational companies are somewhat different. It can be argued that multinationals can well afford the development costs of including smaller languages amongst their multilingual offerings, and that paying for the necessary linguistic resources would be a great help to those languages. However, in the absence of strong legislation requiring Welsh language provision, most multinationals only heed the economic argument, and if the cost of producing or procuring those resources is larger than the anticipated return on their outlay, they will not pay for their development. If, on the other hand, appropriate resources are available to them at no cost, they are then more willing to consider supporting that language amongst their offerings. The minoritized or endangered language community benefits as many of their users already use those products every day in English, Spanish, French or whatever other dominant language they speak.

Additional clarity would however be welcomed in understanding the legal ramifications of different licences as there are many legal grey areas. For example, if new language or acoustic models are trained from a specific corpus, does the licence of the original corpus carry over to the new models? Or when a new MT engine is trained on a certain dataset, how does that affect the licencing of the new product? This is especially problematic when there are different licences for the two languages in a bilingual corpus, especially derived from a translation memory where the copyright of the original language text was not originally made explicit.

If tools and resources are to be shared outside individual projects and institutions, then dissemination is another issue that comes to the fore. International repositories such as Metashare, Github and Docker Hub have made it easier for developers to find resources in different languages, but for the non-expert user, and anyone interested in a specific language, the plethora of different repositories can be confusing. In addition to using the international repositories therefore, a Welsh National Language Technology Portal was established as a ‘one stop shop’ or ‘brochure site’ pointing at the different resources and giving additional guidance and information on their use (Prys and Jones, 2018).

5. Next Steps

When the Welsh government published its Language Technologies Action Plan in 2018, we could see from the account above that some preparatory research and development had already been done. Work had already begun on the three main areas to be addressed: speech technology, machine translation and conversational AI. Further long-term funding is likely to progress in these areas, with a coherent action plan providing further guidance. The five underpinning themes mentioned in the action plan (creating and sustaining digital infrastructure, developing a culture of open innovation, building capacity and digital skills, digital transformation in the public sector and promoting the creation and use of Welsh language digital products and services) are also challenges to be faced. The themes of course are broad in scope and will need cooperation from many different stakeholders. They demonstrate that developing language technologies and using them to revitalize a language cannot happen in isolation from the wider infrastructural and cultural environment.

The research base in Wales remains very small and issues such as building capacity and digital skills need urgent attention. This is being addressed for school children in the new 2022 National Curriculum for Wales (A Guide to Curricul for Wales, 2019), with its emphasis on digital competence as one of the three core essentials (the other two are literacy and numeracy). At the other end of the educational journey there are plans to create a new Masters in Language Technology programme at Bangor University. Delivering new tools and resources for use in commercial products and services is also crucial. There are opportunities here for the emerging creative and software sectors in Wales, with the potential to develop a new domestic market and to venture into wider multilingual markets from a strong bilingual base.

Some of these points were also echoed in a roundtable discussion to promote a strategic vision for Celtic Language Technologies held during the Celtic Congress at Bangor (Prys and Williams, 2019). Many participants were members of the Celtic Language Technologies Group (CLT), a lose grouping of academics who encourage research in language technologies for the various Celtic languages and organise occasional workshops in the field. All six of the modern Celtic languages are minoritized and endangered, with only Welsh and Irish having developed any coherent strategies for using language technologies for language revitalization. During the roundtable discussion the main needs were summarized as follows:

- the sharing of information across researchers working on individual Celtic languages

- working together to improve training and providing courses in Language Technologies
- developing transfer learning methodologies and common language models for our languages
- sustainability and long-term solutions to maintain our resources.

Training and sustainability needs were themes also picked up in the Welsh Government Action Plan and are doubtless also relevant to other minoritized language situations. Improving the sharing of information amongst the wider community and developing joint research proposals for transfer learning and common models are action points for the CLT to take forward. An official planning document on language technology development for the Irish language is also eagerly awaited, and together demonstrate concerted efforts to allow language technologies to contribute significantly to language revitalization, at least in Wales and in Ireland.

6. Conclusions

Welsh is in a fortunate position compared to many other minoritized and endangered languages. The groundwork has been laid for further development of language technologies, and the various stakeholders: academic researchers, public bodies and industry, are poised to take advantage of new opportunities offered by the Welsh Government's Action Plan. Most importantly, the language community itself is engaged, both as consumers of digital materials and devices and as potential producers of new Welsh digital content and software. Language technologies offer a path towards economic regeneration as well as language revitalization, and while there is much hard work still to be done, current progress is encouraging.

7. Acknowledgements

We gratefully acknowledge funding and support from the Welsh Government for the development of Welsh speech and language technologies since 2013.

8. References

A Guide to Curriculum for Wales 2022. 2019. Welsh Government. https://hwb.gov.wales/storage/f8f9760c-64a1-48ea-80fd-db130ad9050b/a-guide-to-curriculum-for-wales-2022.pdf [Accessed: 12 January 2020].

Cooper, S., Jones, D.B. and Prys, D. (2014). Developing further speech recognition resources for Welsh. In John Judge et al., editors, *Proceedings of the First Celtic Language Technology Workshop at the 25th International Conference on Computational Linguistics (COLING 2014)*. Dublin, Ireland. Pages 55-59.

Cooper, S; Jones, D.B & Prys, D. (2019) Crowdsourcing the Paldaruo Speech Corpus of Welsh for Speech Technology. In a special edition on Computational Linguistics for Low Resource Languages, *Information*, 2019. 10, 247.

Jones, D.B. (2015). Docker Container for Moses. Available at https://hub.docker.com/r/techiaith/moses-smt [Accessed: 11 January 2020].

Jones, D.B. (2018). Docker Container for DeepSpeech. Available at https://hub.docker.com/r/techiaith/deepspeech [Accessed :11 January 2020].

Jones, D.B. and Cooper, S. (2016). Building Intelligent Personal Assistants for Speakers of a Lesser-Resourced Language. In Claudia Soria, et al., editors, CCURL Workshop. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resource Association (ELRA). Pages 74-79.

Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; Dyer, C.; Bojar, O.; Constantin, A. and Herbst, E. (2007), Moses: Open Source Toolkit for Statistical Machine Translation. In John A. Carroll; Antal van den Bosch & Annie Zaenen, editors. *'ACL'*, The Association for Computational Linguistics. Pages 177-180.

Mozilla (n.d.). A TensorFlow implementation of Baidu's DeepSpeech architecture. Available at https://github.com/mozilla/DeepSpeech [Accessed: 11 January 2020].

Office of National Statistics. (2012). *Language in England and Wales: 2011*.

Pammi, S., Charfuelan, M., and Schröder, M. (2010). *Multilingual voice creation toolkit for the MARY TTS platform.* In 7th International Conference on Language Re-sources and Evaluation (LREC), Valletta, Malta. Pages 3750–3756.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian,Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý,K. (2011). The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Waikoloa, HI, USA.

Prys, D. and Jones, D.B. (2018 (1)). Gathering Data for Speech Technology in the Welsh Language: A Case Study. In Claudia Soria, et al., editors, CCURL Workshop. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18).* European Language Resource Association (ELRA). Pages 56-61.

Prys, D. and Jones, D.B. (2018 (2)). Language Technologies Portals for LRLs: A Case Study. *Lecture Notes in Artificial Intelligence*. Springer. Pages 420-429.

Prys, D., Prys, G., and Jones, D.B. (2009). *Improved Translation Tools for the Translation Industry in Wales: An Investigation.* Bangor University, Bangor, Wales.

Prys, M. and Jones, D.B. (2019). Embedding English o Welsh MT in a Private Company. In Teressa Lynn et al., editors. *Proceedings of the Celtic Language Technology Workshop.* European Association for Machine Translation. Dublin, Ireland. Pages 41-47.

Prys, D. and Williams, I. (2019). *A Round Table Discussion to Promote a Strategic Vision for Celtic Language Technologies.* Bangor University, Bangor. Available at http://techiaith.bangor.ac.uk/wp-content/uploads/2019/08/A-roundtable-discussion-to-promote-a-strategic-vision-for-Celtic-Language-Technologies.pdf [Accessed: 11 January 2020].

Williams, B., Prys, D. and Ní Chasaide, A. (2005). Experiences of creating a research capability in speech technology for two minority languages. In *Proceeding of the 9th European Conference on Speech Science and Technology (Interspeech)*. Pages 188-191.