National Language Technologies Portals for LRLs: a Case Study

Delyth Prys, Dewi Bryn Jones

Bangor University
College Road, Bangor, Gwynedd, Wales
{d.prys, d.b.jones}@bangor.ac.uk

Abstract

The new Welsh National Language Technologies Portal is an extensive resource for researchers, developers in the ICT and digital media spheres, open source enthusiasts and code clubs who may have limited understanding of language technologies but which nevertheless have a need for incorporating linguistic data and capabilities into their own projects, products, processes and services in order to better serve their wider LRL community. It includes a repository of free, simple and accessible resources with documentation, tutorials, example code and projects. This paper describes the rationale and process of building the Portal, new novel resource dissemination mechanisms employed, such as online APIs and Docker, as well as the lessons learnt and applicability to other similar linguistic situations and communities.

Keywords: LRL, repositories, language resources, Welsh

1. Introduction

One of the many problems facing developers and users of language technology for less-resourced languages (LRLs) is the fragmented nature of resources and attendant information for the languages they work with. In the international field of HLT and computational linguistics, large international catalogues and repositories of language data and processing tools such as ELDA and META-SHARE are now making it much easier to distribute and promote LRs at a global level. However, these are targeted at academics and professionals who are well-versed in HLT, who often work with multiple languages, and don't need a lot of hand-holding to integrate these tools and resources into their finished products or environments.

Such repositories fail to address the needs of many less-resourced languages that historically have been of little interest to the mainstream academic and commercial markets. Researchers and developers struggle to create and disseminate tools and resources to be taken up by local companies and activists who cater for those language communities.

In such scenarios, many of those who service the LRL community are small-scale developers of apps, local media companies, webmasters, open source enthusiasts, coding clubs, and public and third sector bodies. Global companies who integrate a wide range of languages in their multilingual offerings, provided it is not too costly to do so, are also potential users. This wide constituency of developers may vary in its understanding of ICT and digital media in general, but the incorporation of the linguistic element into their products is an additional barrier to the uptake of the necessary tools and resources. In addition to the tools and resources themselves, this constituency therefore has additional requirements for help and support.

In order to help both the local constituency of developers and the LRL community that they serve, it can be beneficial to create local repositories specifically targeting them. Such repositories need to include additional tutorials and case studies on resource use, in addition to the metadata and documentation required of all LR repositories. Resources also need to be accessible, user friendly and simple to use, enabling incorporation into other products with the minimum of additional coding and development.

This paper describes one effort to create a National Language Technologies Portal for one such less-resourced language, namely Welsh, the steps taken to choose the tools and resources for inclusion, the environment created to disseminate language technology code, data and capabilities and the additional tutorials and documentation that accompanied them.

The paper concludes with considerations for the further elaboration of the National Language Technologies Portal, together with lessons that have been learnt that may be applicable to other such LRL communities.

2. Choosing Tools and Resources for Inclusion

The META-NET series of White Papers on Europe's languages in the Digital Age (Rehm and Uszkoreit n.d.) provides an overview of present provision of LRs for many of Europe's languages. Welsh was a late addition to this series (Evas n.d.), and was scored as the weakest of the languages covered in the series in terms of its LRs. Despite this, a significant number of LRs were in existence, many of them not cited in the META-NET volume on Welsh. This in itself was cause for concern, showing low levels of awareness, even within the Welsh national community.

Attempts had been made in the early 2000s to create a Welsh or a Celtic BLARK, emulating the Dutch exercise in creating a comprehensive taxonomy of tools and resources, and elaborating the interdependencies between types of data and applications and modules that use them (Prys, 2006). However, government support was not forthcoming for such a strategic project, possibly highlighting the difference between official European languages and minority languages, even though Welsh was gaining some additional recognition at the time. A strategy document *Information Technology and the Welsh*

Language was published (Welsh Language Board, 2006), but this was a catalogue of various IT considerations, including terminology standardization, localization, bilingual web design, the need for corpora and speech technology, training and second language learners needs, rather than an action plan to provide a comprehensive infrastructure.

Consequently it was left to individual research teams, companies and volunteers to develop any relevant tools and resources, being guided by their own interests and the availability of funding. Many of these LRs were the result of R&D projects at the Language Technologies Unit at Bangor University, a self-financed research centre where the roadmap towards comprehensive coverage of Welsh language LTs had to be balanced by the need to attract funding for its projects. Various tools and resources were created at the LTU for internal project use, for example an online version of a Welsh spelling and grammar checker had been developed in order to gather a corpus of errors and of corrected texts (*Cysill Ar-lein*, 2009). Some of these tools and resources had been made publicly available at various locations, but others that had been intended purely for internal use and were not suitable for public distribution without further refinement and packaging.

The Welsh National Language Technologies Portal was therefore conceived as a project to improve the visibility and dissemination of at least some of the available resources. Funding of £50,000 was obtained from a small Welsh Government fund, with the work having to be completed within one financial year. It was recognised that not all resources could be prepared, packaged and published during the project's initial allotted timeframe of 8 months. However, in order for such a National Language Technologies Portal to be of any worth from the outset, a core set of the right resources needed to be chosen for inclusion.

The project's first task was to create an inventory of available resources and tools, together with notes on their state of readiness, documentation, the opportunities on offer, insights and analytics gathered from any logs or public feedback, licensing issues, target audience and suggested priority (Welsh Government Internal Report, 2014). This differs substantially from a full taxonomy of resources as developed by Binnenpoorte et al (2002) for HLT resources for Dutch, but was an attempt at least to find some thematic groupings from the legacy tools and resources. It was recognised that these were a rather ad hoc legacy collection, including, for example, lexical resources such as dictionaries and applications, and also tools that use these lexical resources, such as a Welsh language detector.

The inventory included over 30 candidates which were grouped into the following categories:

1. Language Proofing
2. Machine Translation
3. Speech Technology
4. Corpus Harvesting and Processing
5. Lexical Resources
6. Welsh Place-names

Each candidate resource was assigned a perceived level of importance within its category to its target audience. The following table examples one candidate resource:

5. Lexical Resources		
Id	**Name**	**Importance**
5.7	Welsh Language Detection	2
Description	A resource that can detect Welsh language texts	
State of Readiness	Developed so far for internal use only. Needs further evaluation and possibly training.	
Documentation	Needs to be documented and packaged with example usages	
Ideas	There are many e.g. separate texts that are linguistically mixed (English and Welsh)	
Comments	The detector is based on an open source Java library. It has been trained with texts from the Language Technologies Unit's various corpus resources. It would be worthwhile to contribute the models back to the original project: http://code.google.com/p/language-detection	
Licensing	Apache	
Target Audience	Software developers, (especially those which process bilingual texts and resources)	
Priority	This is important. This is a missing resource that other developers often contact us about	

Table 1 - Example Candidate Resource Entry

Based on the content and suggestions contained in the report, as well on its own information, the government funders prioritised seven resources and tools for initial inclusion:

1. **Vocab** – a Welsh/English Dictionary Website Plugin that integrates the extensive *Geiriadur Bangor* Welsh-English dictionary into any website
2. **Cysill Ar-lein** - an on-line Welsh language spelling and grammar checker
3. **Welsh/English equivalents word list** – a simple list of corresponding words and phrases in Welsh and English
4. **Welsh Social Media Corpus** – Welsh language texts collected from Twitter and Facebook
5. **Welsh/English Alignment Tool** – based on the popular HunAlign
6. **Welsh Statistical Machine Translation** - data and scripts for facilitating machine translation with the Moses SMT system between Welsh and English
7. **Welsh Language Synthetic Speech** - Welsh text to speech voice based on the Festival Speech Synthesizer system.

The project team undertook that, if time and resources allowed, the following resources should also be added to the initial resources:

8. **Welsh Language Part-of-Speech Tagger** – the PoS tagger used within the Welsh language grammar checker.

9. **Language Detection for Welsh** – as described in Table 1.

This left out many other useful tools and resources, such as a parallel Welsh/English corpus of the Welsh Assembly's Record of Proceedings, despite pointing out its importance for the development of machine translation and other applications. It is however hoped that this and other resources may be added at some future time.

The project proceeded with identifying the common dissemination mechanisms between various resources which would influence National Language Technologies Portal organisation and construction.

3. Portal Construction and Organisation

The resources chosen for initial inclusion differed greatly in form, nature and in their mechanisms for dissemination. Some resources existed only in the form of large data files, such as the social media corpora, whilst others, such as the spelling and grammar checker were functionalities desirable for integration into as many software products and websites as possible. It was realised therefore that no single platform implementation would be able to fulfil all the requirements of a Welsh National Language Technologies Portal. The portal would be constructed as an initial website, serving as a superficial layer of information with resources hosted in reality in a federation of bespoke sub-websites and third party web services that would each best serve each resources' means for dissemination.

3.1 Main Website

WordPress, which is a free and open content management system, suitable for use as a web publishing platform, was chosen as the basis for the development of the initial website. The website would need to be bilingual and with the addition of multilingual plugins, WordPress is especially convenient for the production of bilingual sites, due to the ease of toggling backwards and forwards between two languages when developing new content, and of translating each page. Simple usage of WordPress menus was used to organise and present the resources according to a simplified breakdown of language technologies themes (see Table 2).

Translation	**Speech**	**Cloud**	**Corpora**
Aligning	Text to Speech	API Services	Corpus of Welsh Tweets
Localisation		Widgets	
Machine Translation			Corpus of Welsh Facebook Texts

Table 2 - Initial Menu Structure of the Welsh National Language Technologies Portal

Each choice in the menu would take the user to a simple page containing a description of the theme as well as of each resource contained within, with links to access the true location of the resources. In some cases, links were also included to pages that provide support on how to obtain the resources via the various dissemination mechanisms employed.

3.2 Cloud API Services

The National Portal needed to disseminate resources that were functionalities desirable for integration into as many software products and websites as possible. A bespoke website was constructed for providing access to such language technology capabilities via simple online APIs.

As long as the computer or device has a connection to the internet, these capabilities, such as Welsh language text to speech, would be very easy to integrate into any software capable of communicating via simple HTTP requests and response. Thus complexities and barriers associated with downloading, porting and building code for a given deployment environment are eliminated. In addition, the API approach permits free access to capabilities provided by commercial products not amenable to free distribution of code and data but which could still be incorporated legitimately into other software products and projects, in particular open source.

The initial offering of online API Services provided by the portal were: Cysill Ar-lein Welsh language spelling and grammar checker, Welsh language Parts of Speech Tagger, Vocab, Welsh language Text to Speech Engine, Language Detect and Welsh language lemmatizer.

In order to use the API services, the user must register at the bespoke API Services website after which the user can choose an API, agree to the terms and conditions before receiving a 128-bit integer Globally Unique Identifier (GUID) API key for their use of that API. Terms and conditions can vary between API services. All however aim to prevent misuse and protect the service for all users. Thus users agree to accept a rate limit on the number of requests per hour for every API Service. Terms and conditions for API Services provided by commercial products, request that no attempts be made at reverse engineering.

3.3 GitHub

Some resources exist as code. The National Portal project would also create a significant amount of new code for its tutorial and example projects. The most obvious and most attractive location for hosting open source code freely is GitHub. GitHub is well known and popular amongst developers, where you can discover, use and contribute to millions of projects using a collaborative development workflow.

All code based resources provided by the National Portal would exist within a number of repositories within a specially created GitHub organisation, separate from any other ongoing projects developed on GitHub by the researchers. Repositories were used also to contain documentation, tutorials and example projects for each API service. In hosting resources on GitHub it will be possible for users to contribute additions and enhancements back into the resources.

In addition, the National Portal's main website would provide instructions as to how resources located on GitHub could be downloaded even if Git is not installed on the user's PC.

3.4 Docker

Some of the resources on offer as code and script for download and local execution are of a very complicated and sophisticated nature, none more so than the Moses-SMT machine translation system. It is not a trivial task for a developer, let alone a MT practitioner to master its complicated and lengthy build process and subsequent loading of translation models and or training.

Recent technological developments in software deployment provide opportunities for significantly simplifying the downloading, installation and execution of complex applications to the National Portal's target audience. Docker is the leading solution and service to date in this space. Docker is an open platform for building, shipping and running distributed applications. Entire applications can be packaged as images and executed in containers without worrying about inconsistencies between various development and production environments and without locking into any platform or language.

In addition the Docker Hub Registry provides a free-to-use registry of pre-prepared images submitted by users or officially by popular open source projects such as Ubuntu, Postgresql and WordPress.

Thus, in addition to hosting all Welsh/English Moses-SMT resources on GitHub, an image of a Moses-SMT server packaged with scripts that facilitate fetching pre-trained Welsh/English translation models, and execution were submitted to the Docker Hub Registry.

Thus a user can in two very easy Docker commands have his/her own local Welsh/English Moses-SMT machine translation environment and server.

4 Beta-testing, Engaging and Building a Developers Community

During the period of building the National Portal and packaging the tools and resources, a number of methods were used to reach out to potential users in order to raise awareness of the forthcoming repository and also to find beta-testers who would be prepared to work with us to make sure it was fit for purpose and easy to use.

Since the National Portal's website had been implemented on WordPress, its blogging features were put to use along with our twitter feed to reach out and market resources as they became available to as wide as possible a target audience.

Fortunately also the project was able to reach out and obtain feedback from companies that had previously approached its members in the past enquiring whether such resources as they needed for their products existed for Welsh. The *Hacio'r Iaith* event (an annual Welsh-medium unconference for hackers, enthusiasts, developers and members of the media industry) was also used as an ideal venue to reach a worthwhile target audience.

Amongst the most interesting groups who engaged with beta-testing the resources were a class of 9 and 10 year old school children and their teacher from a small rural school in North West Wales. The school had received a number of Raspberry Pi computers through the generosity of the Raspberry Pi Foundation and Google and wanted to utilise the new Welsh language resources for activities that would teach them not only coding but strengthen their Welsh language written and oral skills. Coding club resources in Welsh are very scarce and thus an existing introduction to coding lesson plans by Raspberry Pi Foundation based on the Turing Test was translated. The lesson plans were adapted to incorporate the Welsh language Text to Speech API service, as well as expanded to suggest using some other language technologies such as the Cysill Ar-lein spelling and grammar checker, parts of speech tagger and language detection, in order to give their Turing test code the appearance of Welsh language capabilities.

Outreach activities climaxed in a one day conference to launch the National Portal and give it further publicity. This was an opportunity to bring together academics, businesses, enthusiasts and other stakeholders, and thereby break down some of the barriers that traditionally exist between them. Presentations included one on the use a freelance developer had made of the Twitter corpus, and another on analysing Welsh tweets, together with the work the schoolchildren in programming their Raspberry Pis to speak Welsh. The conference was covered by the Welsh language television and radio news with the children giving excellent answers in their interviews.

Speakers however were not confined to local participants, with international presenters from Ireland, the Basque Country and South Africa also taking part. This brought the much needed perspective of other less-resourced languages into play, hopefully helping to build up a network of like-minded people, and laying the foundation for future joint projects and collaborations.

5 Further Work and Conclusions

Although this initial project was only 8 months in duration, the National Portal itself was designed for long term sustainability. The repository will continue to be developed and used as and when other projects produce relevant tools and resources.

Already a further project on combining recent developments in Welsh language speech recognition, along with machine translation, text to speech and commercial search APIs by Google and Microsoft called *Seilwaith Cyfathrebu Cymraeg* (Welsh Communications Infrastructure), funded by the Welsh Government and S4C (the Welsh Broadcasting Authority), has promised to make any resources it produces freely available through the National Portal and its dissemination mechanisms be it GitHub, new API Services and/or further images in the Docker Hub Registry.

The popularity of the API Services approach has been encouraging with app developers, webmasters and others normally not able to use language technologies already integrating or at least exploring enhancing their products and digital offerings for Welsh. The Vocab widget and API service has been added to a popular national Welsh language news website with others significant services in the pipeline. The *Cysill Ar-lein* Welsh language spelling and grammar checker API is being considered for inclusion into apps for Welsh learners as well as scripts for automating proof reading Welsh language articles in Wikipedia. In the meantime other developers are asking for further language technology capabilities to be added.

Also encouraging is that the Moses-SMT image in the Docker's Hub Registry has been pulled to date nearly 50

times and has appeared to have created forked versions for other language communities.

The approach of National Portal web sites for other types of Welsh language resources has been already well-established, with a Welsh National Terminology Portal having been set up in 2010, and a Welsh National Corpora Portal following in 2011. Building on the National Portal 'brand' has helped in raising awareness of the new offering of a Welsh National Language Technologies Portal, as well as reflect accurately its national character and provenance of funding.

A common danger with tools and resources for LRLs is lack of quality control, as developers are desperate for anything they can use. This can be overcome not only by the formal evaluation of such tools and resources, but also by encouraging continuous feedback and dialogue with communities of users. The former can prove challenging for LRLs, who may lack the capacity for developing resources, let alone evaluating them, and where methods devised to evaluate LRs for well-resourced languages may not always be suitable. On the other hand, relationships with communities of users, especially if activists and enthusiasts are included, can be closer and better in LRL environments, and feedback and engagement with these communities needs to be actively encouraged, rather than merely aping what works for WRLs.

6 Acknowledgements

The Welsh National Language Technologies Portal project reported on in this paper was made possible with the financial support of the Welsh Government, through its Technology and Digital Media in the Welsh Language Fund. The authors would also like to thank the contributors from various SMEs, hackers and communities of users that assisted the project team.

References

Binnenpoorte, D., F. De Friend, J. Sturm, W. Daelemans, and C. Cucchiarini (2002). *A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch.* In Proceedings LREC 2002, (Third International Conference on Language Resources and Evaluation). Las Palmas de Gran Canaria, Spain.

Cysill ar-lein (2009). Bangor University: Bangor. http://www.cysgliad.com/cysill/arlein/. Accessed 17/09/2015.

Evas, Jeremy (n.d.) *The Welsh Language in the Digital Age.* Metanet White Paper Series: Springer.

Prys, D., Jones, D.B., Cooper, S. and Robertson, P. (2014). *Adnoddau Technolegau Iaith i'w Cynnwys mewn Porth Adnoddau Cenedlaethol.* Unpublished Report to the Welsh Government.

Great Britain (2011). *Welsh Language (Wales) Measure.* London: Her Majesty's Stationery Office. Retrieved from: http://www.legislation.gov.uk/mwa/2011/1/contents/enacted. Access date: September 16, 2015.

Prys, Delyth (2006). *The BLARK Matrix and its relation to the language resources situation for the Celtic languages* in Proceedings LREC 2006, (Fifth International Conference on Language Resources and Evaluation). Genoa, Italy.

Rehm, George and Uszkoreit, Hans (n.d.) White Paper Series. Springer.

Welsh Language Board *(2006). Information Technology and the Welsh Language / Technoleg Gwybodaeth a'r Iaith Gymraeg.* Welsh language version retrieved from http://orca.cf.ac.uk/43799/1/3964.pdf. Access date: October 27, 2015.

Welsh National Corpora Portal (2011). Retrieved from: http://corpws.cymru/?lang=en. Access date: September 9, 2015.

Welsh National Language Technologies Portal (2015). Retrieved from: http://techiaith.cymru/?lang=en. Access date: September 9, 2015.

Welsh National Terminology Portal (2010). Retrieved from: http://termau.cymru/?lang=en. Access date: September 9, 2015.

Raspberry Pi Learning Resources. The Turing Test. Retrieved from: https://www.raspberrypi.org/learning/turing-test-lessons/ Access date: September 9, 2015.