DECHE and the Welsh National Corpus Portal

Delyth Prys
Language Technologies Unit
Canolfan Bedwyr
Bangor University
Wales
d.prys@bangor.ac.uk

Mared Roberts
Language Technologies Unit
Canolfan Bedwyr
Bangor University
Wales
mared.roberts@bangor.ac.uk

Dewi Bryn Jones
Language Technologies Unit
Canolfan Bedwyr
Bangor University
Wales
d.b.jones@bangor.ac.uk

Abstract

This paper describes the on-going project on Digitization, E-publishing and Electronic Corpus (DECHE). It also describes the building of a common infrastructure and portal for displaying and disseminating other Welsh language and bilingual Welsh/English text corpora. An overview is given of other corpora included in the on-line corpus portal, as well as corpora intended for future publication through the portal site. This is done within the context of developing resources frugally and efficiently for less-resourced languages.

1 Introduction

Electronic language corpora are some of the most essential resources both for contemporary linguistic research and the development of new language technology applications. They also present a challenge to Welsh and other Celtic languages as smaller languages that are invariably under-resourced with regards to the availability of and interest in funding language technologies. Existing resources need to be recycled, updated, and presented in accessible formats in order to be useful to a new generation of researchers. Although the whole world wide web is now, in some sense, available as an on-line corpus (Gatto, 2014), and that notable attempts have been made to use it to build linguistic corpora, foremost amongst them Kevin Scannell's Crúbadán Project for less resourced languages (Scannell, 2007), we believe that there is still a need for specific text corpora in different domains and for various uses that are easily searchable and accessible to a wide academic community and beyond. This paper provides a brief overview of how a piecemeal collection of Welsh corpora are being brought together into a coherent, online, freely accessible and expanding Welsh National Corpus Portal (Porth Corpora Cenedlaethol Cymru, [no date])

2. From the first corpus to National Portal

The catalyst for the development of the Welsh National Corpus Portal was the awarding of a grant for the DECHE Project. DECHE (Digido, E-gyhoeddi a Chorpws Electronig, translated as Digitization, E-publishing and Electronic Corpus), is funded by Y Coleg Cymraeg Cenedlaethol (The Welsh National College). This is a virtual college established by the Welsh Government in 2011 to promote and deliver Welsh-medium university education in Wales, including the creation of new Welsh language academic resources. The primary aim of the DECHE project is to produce e-books out of Welsh language scholarly, academic books which are out of print and unlikely to be reprinted in traditional paper format. Candidates for producing as e-books are nominated by lecturers working though the medium of Welsh, and prioritized by the Coleg Cymraeg, according to best fit with the Coleg's Academic Development Plan (Coleg Cymraeg Cenedlaethol, 2011). The current project processes around 30 books a year, which are published on Y Porth (Y Porth, [no date]), the Coleg's own portal website for Welsh

Proceedings of the First Celtic Language Technology Workshop, pages 71–75,
Dublin, Ireland, August 23 2014.

academic teaching resources. Books are scanned at the National Library of Wales, and passed to the project's purpose built OCR software. Human based proofreading and corrections are made before final publishing into E-PUB (readable by most e-readers) and Mobi (for Kindle) formats. PDFs are also produced for the purpose of printing personal copies.

2.1 DECHE Corpus of Welsh Scholarly Writing

The creation of a corpus of academic Welsh writing (named DECHE Corpus of Welsh Scholarly Writing) is a spin off from this primary e-book activity, taking advantage of the fact that these books are being digitized in any case for another purpose. The original OCR process produces a text which still contains many errors, especially in dealing with Welsh accented characters and other linguistic peculiarities. Therefore the human proofreading stage is vital in producing high quality and clean text. Human involvement in the workflow allows in addition for metadata such as the book title, author, date of publication, keywords, and subject fields, as well as limited annotations within the text body to be input into the corpus. To date 30 books have been added into the corpus, giving a total of approximately 450,000 words. This total will rise annually during the lifetime of the project.

2.2 Welsh National Corpus Portal

The Welsh National Corpus Portal was developed as a means of fulfilling not only the secondary objectives of the DECHE project, but also to serve for the first time as an opportunity to plan and present other Welsh language related corpus resources. The corpus portal was inspired by the Welsh National Terminology Portal (Porth Termau Cenedlaethol Cymru, [no date]), which serves as an online one stop shop for displaying and searching tens of standardized terminology dictionaries. Although websites such as that of SketchEngine (Kilgarriff et al, 2004) provide an overarching interface to query many corpora in a number of languages, including Welsh, they deal mainly with major languages with ample corpus resources. Many smaller languages are still poorly endowed with corpus resources of any kind, and the Welsh National Corpus Portal seems a rare example of an attempt to bring together disparate resources for such a language.

The Welsh National Corpus Portal infrastructure supports importing text resources as well as a search tool for both monolingual and bilingual corpora. A corpus management interface was developed in-house in order to facilitate tasks such as importing texts to the on-line corpora, using a 'submit to website' button by project staff, without the intervention of software experts. In the case of the DECHE corpus, the infrastructure supports importing the finally published e-pub files. Publication level metadata is also collected and stored in the infrastructure with imported texts. In the case of bilingual corpora, CSV and TMX file formats are supported.

The corpus portal's search and import functionalities employ natural language processing components for segmentation and lemmatization. The segmentation tool was originally developed in-house for use in translation memory software, and the lemmatizer was originally developed for the Cysill grammar and spelling program. Lemmatization enables searching through the Portal's Welsh language texts for all forms of a given search word, including mutations forms and conjugated verbs. For example, typing in 'canu' (to sing) will also return all possible forms of the lemma, including 'cenir' (present impersonal form of the verb), 'ganodd' third person past tense with soft mutation, 'nghanu' (verb noun form with a nasal mutation) and 'cheni' (second person present form of the verb with a spirant mutation).

3 Other published corpora

3.1 Criteria for including corpora in the National Portal

To date, only corpora developed at the LTU itself, or that the LTU has inherited responsibility for, are included in the Portal. This is for practical reasons, in that these corpora have been designed or adapted in house specifically for inclusion in the Portal, their format is compatible with that of the Portal.

Other useful Welsh corpora available on-line are listed on the web-site, with a link provided to their own web-sites. It is hoped in future that corpora from other sources will become available for inclusion in the National Portal, and that information about other unpublished corpora will also be made available there.

3.2 The CEG corpus

The first major Welsh electronic corpus to be collected was the CEG (Corpws Electroneg o'r Gymraeg) corpus in the early 1990s (Ellis, N.C. et al. 2001). This was designed as a lexical and word count corpus of samples of around 2,000 word segments from various genres of fiction and non-fiction resulting in a 1 million word corpus. This was innovative in its time and together with an associated part of speech tagger and lemmatizer, was a major contribution to Welsh corpus studies. However, as time went by CEG became difficult to access and use, and due to numerous requests for help from individual scholars, the decision was made to port it, together with the attendant metadata, into the Welsh National Corpus Portal. The original CEG files and data however were also ported to a new server and have been maintained on-line in addition to the new format.

3.3 The National Assembly for Wales Record parallel text corpora

Similar to the Hansard produced by the UK parliament at Westminster, the National Assembly of Wales produce and publish a bilingual record of its main chamber's proceedings. Assembly members may speak in either Welsh or English. Their words are transcribed and translated into the other language, creating a bilingual record of what is said. The written proceedings are carefully translated and edited, and thus provide an excellent resource for a variety of research and development purposes.

An early version of a parallel text corpus from the National Assembly of Wales Record was created by Jones and Eisele in 2006 (Jones et al, 2006). This covered the period of the first assembly 1999-2003 and has been included into the Welsh National Corpus Portal. A further corpus produced by the CATCymru project (CATCymru, 2009), covering the third assembly (2007-2010) has also been included into the Portal. Both corpora provide in total approximately 850,000 parallel segments and a word count of 20 million. Thus when added together these corpora have been a valuable resource for a wide spectrum of users from statistical machine translation practitioners to freelance translators who make heavy use of the portal's search facilities in conjunction with their terminology searches.

The National Assembly for Wales has streamlined and simplified its publication of the record. It also currently has a machine translation strategy with Microsoft to speed up the translation of the Assembly Record and lower costs. It is likely therefore that this particular collection in the corpus portal will grow substantially over the coming months and years.

3.4 The experimental language register corpus

This is a very small corpus extracted from the much larger but as yet unpublished Corpws Cysill Ar-lein (see 4.1). Its purpose is to study linguistic features of various language registers, especially with a view to developing methods of accurately tagging and recognizing texts according to their language register. This forms part of a Welsh Government and S4C project on speech recognition, but it is foreseen that a corpus of language registers will also be of much wider interest to the academic community. The corpus is still under development, but can already be accessed through the Welsh National Corpus Portal.

4 Unpublished corpora

4.1 The Cysill Ar-lein corpus

A special, free on-line version of a Welsh language spelling and grammar checker, Cysill (Cysill Ar-lein, 2009), was created with a view of collecting user generated samples of Welsh texts. This automatically generates a corpus of errors, with the corrected texts collected also from the users'

sessions. The on-line version of Cysill has been very popular for a number of years. During four months of use in early 2014, the corpus grew by 2.5 million words, an average monthly total of 650,000 words of text throughput. To date the corpus comprises upwards of 14 million words each in corrected and uncorrected versions.

An analysis of the content shows a wide variety of text types, ranging from school and student essays to job applications, journalistic articles, formal documents, blogs, tweets and e-mails. Although use of this corpus for academic research was clearly stated in the terms and conditions, with warnings concerning privacy and confidentiality, sensitive material such as job applications and CVs containing names and addresses are common in it. Publication has been frustrated by users' lack of attention to these warnings, and it is inadvisable to publish without reasonable quality of anonymization. It is however available internally for research purposes, and has been used by staff and students, notably by Wooldridge (Wooldridge, 2011) in her MRes study of interference from English on Welsh texts.

4.2 The corpora of 19th century and World War I Welsh newspapers

The latter part of the nineteenth century and beginning of the twentieth century was the golden age of Welsh newspaper publishing. There was a large literate Welsh public who had not yet learnt English who supported a thriving Welsh language press. A recent project to create a website of resources for the First World War in Wales (Cymru 1914, [no date]), sponsored by JISC (a registered charity that champions the use of digital technologies in UK education and research), and led by the National Library of Wales, included a task to provide gist machine translation of the Welsh language newspapers into English. Unlike translations carried out to an accepted standard by human translators, gist translations only aim to provide a rough idea of the contents. They need not be polished in terms of language or always accurate in terms of meaning, but they provide a quick and cheap way to access source texts in a language which is unknown to the reader.

In order to complete this task, digitized copies of the Welsh newspapers from the war period were used, totalling approximately 11 million words. The much larger collection of digitized pre-war collection of Welsh newspapers, totalling approximately 223 million words was also received by the project, to be used as training data. Both these bodies of texts were imported into the Welsh National Corpus Portal infrastructure for ease of manipulation.

The quality of the digitization of these newspapers is not very high, due to the poor ink and paper quality. The unstandardized orthography and old fashioned language can also cause difficulties, and further work on this corpus could include the use of automatic standardization techniques similar to those used by Scannell (Scannell, 2009) for the Irish language. Nevertheless this could still be a very valuable corpus of Welsh, especially since it is by far the largest corpus of Welsh available. Efforts are currently under way to obtain permission to publish these corpora. In the meantime they are searchable internally and have been used in a chunking exercise to develop language models for speech technology and machine translation for Welsh.

6. Conclusion

The Welsh National Corpus Portal to date includes a corpus of contemporary academic Welsh, a legacy corpus of Welsh designed for word count and lexical purposes, a bilingual corpus of parliamentary Welsh, and an experimental corpus of different language registers. Adding to these the Cysill corpus of errors, and the nineteenth and early twentieth century newspaper corpora gives us an unexpectedly broad and deep range of Welsh language corpora. Given that only the DECHE corpus and experimental corpus of registers received any grant funding, and only as secondary considerations to other primary objectives, a great deal has been accomplished in recent years. Further work aims to expand the collection and integrate more natural Welsh language processing tools to aid annotation analysing and searching. It is hoped that the Welsh National Corpus Portal will continue to grow and provide inspiration for other less-resourced languages facing similar challenges.

References

Adam Kilgarrif, P. Rychly, P. Smrz, D. Tugwell. 2004. The Sketch Engine. *Proc EURALEX 2004, Lorient, France Pp. 105-116*. http://www.sketchengine.co.uk

CATCymru. 2009. *Cyfieithu â Chymorth Cyfrifiadur: Computer Assisted Translation.* [Online] Available at: http://www.catcymru.org/wordpress/?p=13043. [Accessed: 8 May 2014].

Coleg Cymraeg Cenedlaethol. 2011. *Coleg Cymraeg Cenedlaethol Academic Plan.* Available at: http://www.colegcymraeg.ac.uk/en/media/main/dogfennau-ccc/dogfennaucorfforaethol/CCCAcademicPlan.pdf [Accessed: 8 May 2014].

Cymru 1914. [No date]. *The Welsh Experience of the First World War* [Online] Available at: http://www.cymru1914.org [Accessed: 8 May 2014].

Cysill Ar-lein. 2009. *Welsh Spelling and Grammar Checker* [Online] Available at: http://www.cysgliad.com/cysill/arlein/ [Accessed: 8 May 2014].

Dafydd Jones & Andreas Eisele. 2006. Phrase-based Statistical Machine Translation between English and Welsh. *LREC Conference Proceedings.* Available at: http://www.mt-archive.info/LREC-2006-Jones.pdf [Accessed: 8 May 2014].

Dawn Wooldridge. 2011. *Gwella Cysill at Ddefnydd Cyfieithwyr: adnabod ymyrraeth agan yr iaith Saesneg mewn testunau Cymraeg.* MRes, Bangor University. Available at: http://www.cyfieithwyrcymru.org.uk/adnoddau-4.aspx [Accessed: 18 June 2014].

Kevin P. Scannell. 2007. *The Crúbadán Project: Corpus building for under-resourced languages.* [Online] Available at: http://borel.slu.edu/pub/wac3.pdf [Accessed: 9 May 2014].

Kevin P. Scannell. 2009. *Standardization of corpus texts for the New English-Irish Dictionary*. Paper presented at the 15th annual NAACLT conference, New York, 22 14 May 2009. Available at: http://borel.slu.edu/pub/naaclt09.pdf [Accessed: 18 June 2014].

Maristella Gatto. 2014. *Web as Corpus Theory and Practice.* Bloomsbury Academic.

N. C. Ellis, C. O'Dochartaigh, W. Hicks, M. Morgan, & N. Laporte. 2001. *Cronfa Electroneg o Gymraeg (CEG): A 1 million word lexical database and frequency count for Welsh.* [Online] Available at: http://www.bangor.ac.uk/canolfanbedwyr/ceg.php.en [Accessed: 8 May 2014].

Porth Termau Cenedlaethol Cymru. [No date]. *Welsh National Terminology Portal* [Online]. Available at: http://termau.org/?lang=en [Accessed: 8 May 2014].

Porth Corpora Cenedlaethol Cymru. [No date]. *Welsh National Corpus Portal* [Online]. Available at: http://www.corpws.org/?lang=en/ [Accessed: 8 May 2014].

Y Porth. [No date]. *Y Porth* [Online] Available at: https://www.porth.ac.uk/en/ [Accessed: 8 May 2014].