Using CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – National Corpus of Contemporary Welsh

Dawn Knight, Steve Morris, Tess Fitzpatrick, Jonathan Morris, Paul Rayson, Irena Spasić, Enlli Môn Thomas, Alex Lovell, Ignatius Ezeani, Steven Neale, Jennifer Needs, Scott Piao, Mair Rees, Lowri Williams, Laura Arman, Vigneshwaran Muralidaran, Bethan Tovey, Laurence Anthony, Thomas Michael Cobb, Margaret Deuchar, Kevin Donnelly, Michael McCarthy and Kevin Scannell.

This half-day workshop provides a working demonstration of the ESRC/AHRC-funded CorCenCC (Corpws Cenedlaethol Cymraeg Cyfoes – National Corpus of Contemporary Welsh – see www.corcencc.org) corpus. CorCenCC is the first large-scale general corpus of Welsh language, comprising 10 million words of spoken, written and electronic language, and has adopted a user-based design, informed, from the outset, by representatives of all anticipated academic and community user groups.

The 42-month CorCenCC project will finish in 2019, so this workshop will allow provide an opportunity for academics and non-academics alike to see the near-final outputs of the project and stakeholders to provide feedback on the tools that we have been developing. CorCenCC will generate demonstrable impact by enabling community users (including those in the public/private or third sectors) to investigate dialect variation or idiosyncrasies of their own language use, professional users to profile texts for readability or develop digital language tools; teachers to develop and utilise pedagogic tools and materials, and researchers to investigate patterns of language use and change.

The CorCenCC project has involved 1 Principle Investigator (PI), 7 Co-Investigators (Co-Is) from 4 top-ranked academic institutions (Cardiff, Swansea, Bangor and Lancaster); 5 RAs; 2 PhD students; 6 consultants, and 9 partners/stakeholders from a range of disciplines, institutions and professional backgrounds including BBC Cymru, S4C, Welsh Government, National Assembly for Wales, WJEC, Welsh for Adults and Gwasg y Lolfa, to help to widen the potential impact of the work, and we intend to invite as many of these stakeholders to this workshop co-located with the Corpus Linguistics 2019 conference.

This half-day workshop brings together members of the team from all work packages on the project, combining poster presentations on issues related to the design and construction of the corpus, with detailed walk-through demonstrations of the key functionalities of the corpus. The workshop aims to target future users of the corpus (from a range of different groups, with a range of interests and expertise), giving attendees a hands-on experience of using CorCenCC, helping to both underline the utility of the corpus and the multitude of ways it can be used to explore questions about contemporary Welsh language use, and the ways in which CorCenCC can be used to facilitate Welsh language teaching and learning at different ages and levels (from beginners to more advanced).

This half-day workshop complements the more detailed research presentations submitted by the project team to the main conference and includes the following:

1. **Context, design and data [poster presentation]**: this focuses on data, presenting the key issues and challenges faced when collecting, collating, transcribing, quality controlling and anonymising corpus data. This includes an overview of the design frame used in the construction of CorCenCC; a demonstration of (and reflections on) the novel CorCenCC crowdsourcing data collection app, and a demonstration of a streamlined, searchable database system for recording information about data collection and metadata. Some guidelines for good practice for corpus construction of
minoritised languages will also be provided here, along with examples of the sorts of questions about Welsh language use that may be answered using this corpus.

2. **Taggers [poster presentations]:** these posters focus on the design, construction and evaluation of taggers. Here, we will provide a working demonstration of both the recently released CorCenCC POS (part-of-speech) tagger, *CyTag*, and the Welsh-language adaptation of UCREL’s Semantic Analysis System (USAS) tagger, *CySemTagger*. Details of how *CyTag* and its bespoke tagset have been integrated into the semantic tagger, and how we have developed larger semantic lexicons to increase the lexical coverage of the semantic tagger will be provided.

3. **Searching and using CorCenCC [hands-on demonstration]:** this part of the workshop focuses on CorCenCC’s infrastructure, outlining to workshop attendees how the corpus can be analysed to explore patterns in contemporary Welsh language usage. CorCenCC has a web-based interface for accessing the tools, similar to online tools such as the BNCweb, Corpus of Contemporary American English (COCA), the Michigan Corpus of Academic Spoken English (MICASE), and the English Native Edited Japanese Essays (ENEJE) corpus. Although there are some existing web-based tools available for hosting corpus data online we developed our own bespoke tools to allow us to have more control over the interface; to tailor functionalities to the CorCenCC metadata and POS and semantic tagsets, and to allow us to integrate seamlessly with the pedagogic toolkit.

4. **Pedagogic toolkit [hands-on demonstration]:** a key innovation of CorCenCC is that it is the first corpus that is integrated with an online Welsh-based pedagogical toolkit which works directly with the corpus data to support language learning and teaching. Inspired by the online Compleat Lexical Tutor data-driven language learning toolkit ([https://www.lextutor.ca/](https://www.lextutor.ca/)), the CorCenCC toolkit includes a Vocabulary Profiler, frequency-based Cloze (gap-fill) exercise builder, reading comprehension assistant and a user-built lexicon with quiz builder. The final part of the workshop, therefore, provides a detailed overview of these functionalities in use, and demonstrates how they may be practically implemented in educational contexts to facilitate the teaching and learning of Welsh to students of different ages and different levels of ability.

For the CorCenCC workshop, we will require a computer lab for participants to engage in hands-on access to the corpus and software tools, plus space for three posters on the wall or poster boards.

The research on which this workshop is based is funded by the UK Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC) as part of the *Corps Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh): A community driven approach to linguistic corpus construction* project (Grant Number ES/M011348/1).