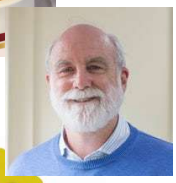
Corpws Cenedlaethol Cymraeg Cyfoes

National Corpus of Contemporary Welsh



Contents

Greetings from the PI

Welcome to the September edition of the CorCenCC newsletter! The summer is (sadly) over and we move into the new academic year and the final 24 months of the CorCenCC project. The summer months have involved some intensive spoken data collection by the team, and dissemination of the work from the project at a range of academic conferences. The first week of September also saw all members of the CPT team (the PI, Co-Investigators and RAs) travel from Bangor, Lancaster and Swansea to Cardiff for our 6-monthly catch-up meeting. This meeting was multifunctional: it gave the core team some much needed time to catch-up face-to-face, to take stock and positively reflect on the great progress we have made so far, and to prioritise and plan the next few months on the project. It was a really productive meeting and, as ever, great to catch up with everyone again! Over the next few editions of the newsletter we will provide you with specific updates from the WPs (starting with WPs 1, 3 and 4) as well as details of our vision and plans for the final 24 months of the project and, as ever, more details of how you can get more involved!

Happy reading! Dr Dawn Knight

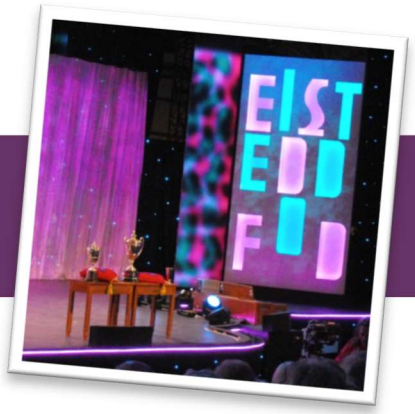
Siarad Cymraeg?

If so, then come along to Tŷ'r Gwrhyd, Pontardawe on 17 November between 13:00 and 15:00 and become a film critic for the afternoon! We'll be showing short clips from different videos and letting you and your friends tell us what you think of them. Those attending the event (which is part of Swansea's programme for the annual Being Human festival) will get the chance to hear more about the CorCenCC project... and also how to download and use the app. Come along, give us your Welsh and meet Steve Morris and Tess Fitzpatrick from the team to find out more.

+ News and events

In August, three of the CorCenCC research assistants travelled to Anglesey to visit the National Eisteddfod. As one of Wales's most important festivals, where every performance, competition and ceremony happens through the medium of Welsh, this was a fantastic opportunity for us to meet a large number of Welsh speakers and encourage them to contribute to our project. The CorCenCC project aims to reflect the wide range of different ways that Welsh is used today. So it's really important to include elements that are unique to Welsh culture, such as people using Welsh at the Eisteddfod. We collected examples of people adjudicating competitions, giving lectures or presentations, and – of course – enjoying themselves over a pint, a cuppa, or an ice cream! Having said that, it wasn't ice cream weather every day, unfortunately. With the heavy rain and strong winds it was challenging to find places suitable to record … and sometimes any people to record! But the sun did come out too, and it was a pleasure to meet Welsh speakers who had come from all over Wales to celebrate their language and culture. Many thanks to everyone who agreed to being recorded – your contributions have helped us make sure that the corpus really includes a bit of everything! But the Eisteddfod is not just a festival for fluent Welsh speakers. One of the wonderful things about it is the fact that it offers an excellent opportunity to people who are learning Welsh. With such a Welsh atmosphere, the Eisteddfod is an ideal place for learners to hear and practise Welsh, and a series of special competitions is available for learners of every level. The Eisteddfod is a travelling festival, and in 2018 it will come to the 'home' of CorCenCC, i.e. Cardiff, so you can be sure that there will be a crew of us there representing the project! By then, part of the corpus will have been built and you will be able to play around with some of the tools we will have developed – so do come along and meet us!

https://eisteddfod.wales/



BAAL 2017 Conference, Leeds University 31/8-2/9

BAAL is a professional association with an international membership of just under 1000 members, which provides a forum for people interested in language and applied linguistics. For more information visit: www.baal.org.uk

The CorCenCC Management Team (CMT - Dawn Knight, Steve Morris and Tess Fitzpatrick) were pleased to have the opportunity to present a poster and a talk at the BAAL 2017 conference in Leeds, at the beginning of September. Our poster outlined the work Steve and Tess had done previously on creating pedagogical wordlists in Welsh using techniques inspired by Français Fondamental (Gougenheim et al, 1964; Morris 2011) and by word association research. CorCenCC offers a frequency-driven way of compiling wordlists, and comparisons with those earlier lists will prompt new research questions! Dawn and Steve presented our paper about the aims of the project, and our progress so far. The audience included applied linguists working on under-resourced languages as well as those interested in corpus linguistics, and we were encouraged by the number of comments relating to the capacity for CorCenCC to act as a blueprint for corpus work in a range of other languages.

CorCenCC on the road

27/9

One important element of the CorCenCC project is 'engagement' and on a very wet night in September, Steve Morris was invited to address Cymdeithas y Llan a'r Bryn in Llangennech, Carmarthenshire. The audience showed great interest in the project and there was quite a lot of discussion about including 'vulgar language/slang' in the corpus and what exactly 'correct Welsh' means. A number of those present agreed to contribute their Welsh to the project (this is the area of the famous "*Shwt ife heddi?*" expression) and we will be recording a whole session from the Society's 2017-18 programme later on in the year. As the project grows and develops, our intention is to take CorCenCC out '*on the road*' more often in order to engage with the wider Welsh-speaking community. If you have any suggestions, please get in touch!

Photo: Glendon Davies (Treasurer), Steve Morris, Colin Lee (Programme Organiser)

Work Package (WP) 1, 3, 4 updates

WP1:

Collect, transcribe and anonymise the data (lead: Steve Morris)

As you will have seen in our previous newsletters, we will be collecting examples of spoken, written and electronic Welsh to include in the corpus. Our main priority at the moment is to collect the spoken examples, since these require an extra level of processing – namely, transcribing. Our transcribers are working very hard, and we really appreciate their help in preparing the corpus content. Many many thanks!

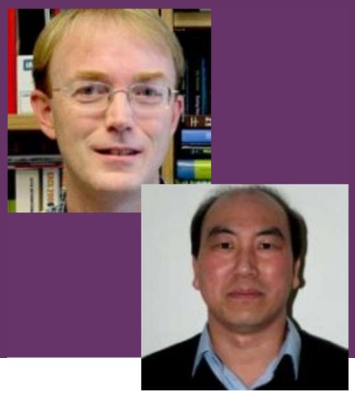
If you would like to join our team of transcribers, please email trawsgrifio@corcencc.org for further details.

Whilst the transcribers have been working on creating a written (typed) record of the content of the audio recordings, the research assistants have been busy collecting more recordings! Since our last update, we have recorded Welsh speakers in Gwynedd, Ceredigion and South Wales, and at the Urdd Eisteddfod and National Eisteddfod. It is lovely to be able to get to know different areas, and we will be visiting every county in Wales by the end of the data collection period. Please email corcencc@cardiff.ac.uk if you would like to know when we will be in your area. We would like to meet as many of you as possible, so as to truly reflect how Welsh language use varies from region to region, from context to context, and from person to person.

We have also begun collecting examples of media language, by recording excerpts of S4C television programmes. Before long, we hope to start recording excerpts of BBC Radio Cymru programmes too. Welsh is fortunate to have special radio and television channels, and CorCenCC is very fortunate that S4C and the BBC have agreed to collaborate with us, to help us create a 'snapshot' of contemporary Welsh in Wales.

WP3:

Develop a semantic tagger for Welsh and semantically tag all data (lead: Paul Rayson)



In WP3, we have been continuing the research on the development of an efficient Welsh semantic tagger, including the building of larger semantic lexicons to increase the lexical coverage of the tagger, and to introduce more effective methods for identifying correct word meanings, and make the tool conveniently accessible by users. In the Lancaster UCREL summer school in June 2017, the current version of a semantic tagger was introduced and promoted as part of the UCREL multilingual semantic tagging system to corpus linguists and natural language processing researchers, providing access to the system via both web service API and a desktop application (a Graphical User Interface for the Welsh tagger web service is downloadable from http://ucrel.lancs.ac.uk/usas/gui/). Although it is still in prototype stage,

it already provides a useful tool for extracting and analysing semantic information from Welsh language data.

The next major steps include: 1) integrate the new Welsh Part-of-Speech (POS) tagger CyTag under development in the CorCenCC Project, which employs a newly designed fine-grained POS tagset; 2) based on the gold standard manually annotated corpus that is under construction, we will evaluate and integrate more context aware word sense disambiguation methods; 3) more Welsh words and multiword expressions will be semantically classified and integrated into the semantic lexicons by detecting synonymous words and expressions. The continuously improved semantic tagger will be demonstrated in various corpus linguistics and natural language processing events in the coming year, and will start to be applied to the CorCenCC corpus data under compilation.

WP4:

Scope/construct the online pedagogic toolkit (leads: Enlli Thomas and Tess Fitzpatrick)

Since our last update (May, issue 11), we have been working with members of the Work Package 5 team to design the pedagogic toolkit. The WP5 team will be putting the WP4 team's ideas into action technologically to create the online resource, so close collaboration is very important to ensure that our plans are feasible.

We are also continuing to consult Welsh teachers and learners, so that the design of the pedagogic toolkit is informed by their needs. However, there are many changes on the horizon in terms of needs and expectations in the world of learning Welsh, with new curricula being developed for adult learners and schools alike. In September, Enlli Thomas, the WP4 lead, was invited to present the concept behind the pedagogic toolkit to the group responsible for developing the Languages, Literacy and Communication area of the new curriculum currently being developed for Welsh schools. This group of around 40 individuals comprised representatives from all over Wales, including a cross-section of teachers from Pioneer Schools representing various Key Stages of learning, bilingual, Welsh-medium and English-medium education, and mainstream and ALN provision, as well as Estyn inspectors and Welsh Government officials from the Curriculum and Assessment Division.

The discussions we had were extremely useful, and the interchange of ideas that happened over the course of the morning enhanced our thinking about how to steer the work as we move forward. The group has agreed to invite us back later on to update them on our progress, and are happy to provide feedback on what we develop.

Thank you very much!

+ Did you know?

The pedagogic facilities will include 'language awareness raising' tools which will function to enable and encourage learners (L1 and L2) to engage with language patterns in order to facilitate learning. These will help to generate quizzes and tasks for learners based on corpus evidence.

+ Meet the team:

Professor Colin Williams, Project Advisory Group (PAG) Member

I am a Social Scientist with a training in Geography and Politics and have been a Professor of Geography at Staffordshire University, a Fulbright Professor of Geography at Pennsylvania State University and I remain an Adjunct Professor of Geography at the University of Western Ontario.

Currently I am based at St Edmund's College, Cambridge where I work on aspects of Post-Conflict Reconstruction and on Multilingual Policy. Prior to my move to Cambridge in 2015 I was a Research Professor in the School of Welsh, Cardiff University for twenty-one years where I remain an Honorary Professor. My field of interest at Cardiff is language policy and planning and for over a decade I combined my university teaching and research interests with being a member of the Welsh Language Board where I specialised in strategy, innovation, government intervention and comparative European and Canadian policy. I continue to advise the Welsh Government on matters of policy development.

I am glad to be a member of the CorCenCC project and I appreciate the team's diverse and very specialist range of disciplines and skills. When fully harnessed the outputs of the team's collective work should provide an excellent resource in the medium term to assist in the capacity-building efforts of academics, public servants, practitioners and citizens to better serve the bilingual needs of a multilingual Wales.

For me the most urgent part of the team's work is to convince civil society to make the most of the results of the constructed Corpus, whilst the most exciting part is the realisation that in the long-term many of the CorCenCC strands will contribute in ways hitherto unimagined at present. Just as with other major databases in the past, such as the Dictionary of the Welsh Language (GPC), the visionary perspective of the CorCenCC founders will provide a ready time-series resource which I hope will be used in at least three ways, namely 1) as a source of knowledge and data which reflects the vitality and diversity of the manner in which Welsh is used currently; 2) as a tool for the refinement of specific policy interventions and programmes in educational practice, community development, the media and IT developments; and 3) as a fundamental platform for the ongoing advancement of Welsh as a language of employment, commerce and professional life so as to realise the full potential of working within a bilingual society in a multilingual context.

My role in CorCenCC is to advise on strategic implementation and the project's liaison with governmental agencies in Wales and beyond.

+

'When fully harnessed the outputs of the team's collective work should provide an excellent resource in the medium term to assist in the capacity-building efforts of academics, public servants, practitioners and citizens to better serve the bilingual needs of a multilingual Wales.'

- Professor Colin Williams.

And finally…

As noted in our last newsletter, Gareth Watkins is leaving us at the end of September. We would all like to wish you well, Gareth, as you pursue a new direction in Mold and thanks once again for your work and your friendship over the last 19 months since the project began. Very best wishes to you from all the team: don't be a stranger!

Contact us

You can keep up to date with developments on the project via Facebook www.facebook.com/CorCenCC/; Twitter https://twitter.com/corcencc (Tweet us @CorCenCC). You can also contact us on the project email address: corcencc@cardifff.ac.uk or visit our website at: www.corcencc.org

CorCenCC is an ESRC/AHRC funded research project (Grant Number ES/M011348/1). The CorCenCC team includes **PI** - Dawn Knight; **CIs** - Tess Fitzpatrick, Steve Morris, Irena Spasić, Paul Rayson, Enlli Thomas, Mark Stonelake and Jeremy Evas; **RAs** - Steven Neale, Jennifer Needs, Mair Rees, Scott Piao and Gareth Watkins; the **PhD student** - Vigneshwaran Muralidaran; **Consultants** - Kevin Donnelly, Kevin Scannell, Laurence Anthony, Tom Cobb, Michael McCarthy and Margaret Deuchar; **Project Advisory Group** - Colin Williams, Karen Corrigan, Lliion Jones, Maggie Tallerman, Mair Parry-Jones, Gwen Awbery, Emyr Davies (CBAC-WJEC), Gareth Morlais (Welsh Government), Owain Roberts (National Library of Wales), Aran Jones (Saysomethingin.com) and Andrew Hawke (University of Wales Dictionary of the Welsh Language). If you have any comments or questions about the content of this newsletter please contact Dr Dawn Knight: KnightD5@cardiff.ac.uk