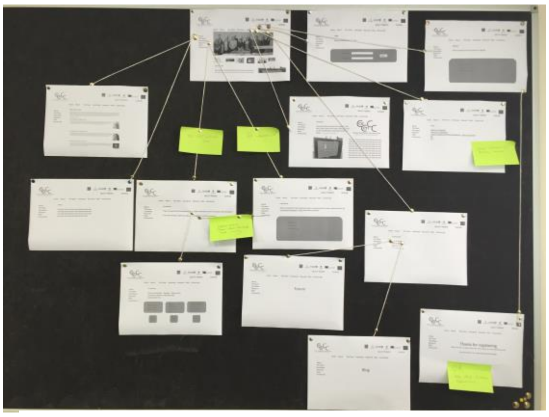
CorCenCC Newsletter

Issue 5: August 2016

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

Welcome to the 5th edition of the CorCenCC newsletter. How time flies when you are having fun! The majority of this summer (so far) has been spent making headway on tasks related to individual project work packages (WPs – updates included below) as well as spreading the word about the CorCenCC work at various academic conferences and public events. I am very happy to say that we already have over 100 people who have subscribed to the newsletter and/or are signed up to contribute data to the project – and our recruitment drive has only just begun! We are also in the process of putting together the official CorCenCC project websites (www.corcencc.org and www.corcencc.cymru) which, we hope, will not only help to build and strengthen the CorCenCC community even further, but will also springboard the data collection phase of the project. We are hoping to launch the sites by the end of September at the latest – so keep your eyes peeled for updates/more information on this.



Storyboarding the CorCenCC websites

Back to the current edition: this month we bring you the latest CorCenCC team news; write-ups from recent conferences and events; WP updates, and we also introduce you to another member of the CorCenCC project team, Mark Stonelake, a Co-Investigator (CI) based at Swansea University.

Happy reading, Dr Dawn Knight (Cardiff University)

News

We want to take this opportunity to introduce you to the newest member of the CorCenCC team, Lorena Varona. Lorena will be working as a part-time Administrative Assistant for the duration of the CorCenCC project and will be based at Cardiff University. Lorena commented that *'a better understanding of our language is essential to know about our culture and to preserve our roots. I'm really pleased to be part of the CorCenCC project team and help to find out more about the use of the Welsh language. Diolch!'* Welcome to the team, Lorena!

Conferences and events

The National Centre for Learning Welsh conference – 8th July 2016

Project lead, Dawn Knight, was invited to give a talk on CorCenCC at the inaugural National Centre for Learning Welsh conference, held at the Marriott hotel in Cardiff on 8th July 2016. The talk, entitled 'Corpora and Pedagogy: developing the community-driven National Corpus of Contemporary Welsh', aimed to inform delegates of the general project plans and to outline how the corpus will be of use to them. We look forward to engaging with the Centre throughout the project, and to updating them on our progress. Many thanks to Efa Gruffudd Jones (Chief Executive of the Centre) and Helen Prosser (Director of Strategy) for inviting Dawn to speak at this event!

WISERD – 13th-14th July, Swansea University

Catching some rays over the Bay

The sun was certainly shining as the CorCenCC Management Team (Dawn Knight, Tess Fitzpatrick and Steve Morris) travelled to Swansea University's new Bay Campus to deliver a paper entitled 'CorCenCC – Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh)' at the annual WISERD conference. WISERD, the Wales Institute of Social and Economic Research, Data and Methods, is based within five Universities in Wales: Aberystwyth, Bangor, Cardiff, South Wales and Swansea, and works in partnership with all UK Universities. The annual conference is Wales' largest social science conference, making it the perfect platform from which to discuss the social, economic and educational impact that the CorCenCC project is likely to have.

National Eisteddfod, Abergavenny

During the first week of August, members of the CorCenCC team visited Abergavenny, home of the 2016 National Eisteddfod of Wales. CorCenCC Co-investigator, Steve Morris, and two Research Assistants, Gareth Watkins and Mair Rees, gave presentations in the Cardiff and Swansea University tents, to introduce the CorCenCC concept and outline our progress thus far. One of CorCenCC's ambassadors, Nia Parry, added to the presentations, speaking passionately about how innovative the project is and how beneficial it will be for Welsh speakers and learners alike.

In the coming months, we will be ready to begin collecting language data for the corpus, and a very important element will be conversations and texts by members of the public. The Eisteddfod provided a perfect opportunity for the CorCenCC project to reach a wide audience of Welsh speakers, so Research Associate, Steve Neale, and Research Assistant, Jennifer Needs, joined Gareth and Mair, in order to distribute information and generate interest among Eisteddfod-goers.

This was Steve Neale's first experience of the Eisteddfod, and he summarises our Eisteddfod visit best: *From the perspective of a learner of Welsh – and one in the very early stages of study – the Eisteddfod was a hugely welcoming place to be. The pride with which people use and promote their language is clear for all to see, but so is the encouragement they will show to those who are interested in exploring the culture of Wales and its language. People's reactions to the CorCenCC project reflected this sentiment – there is a genuine and obvious interest in tools and ideas that can make the Welsh language accessible to as many people as possible.*

Next year, we will be at the Eisteddfod once again – this time to show how you can contribute to the corpus via an app on your mobile phone. So if you missed us in Abergavenny, come to see us in Anglesey in 2017!

Updates from CorCenCC Work Packages (WPs 1-5)

Work on the project is distributed across 6 coordinated work packages, each with specific tasks, aims and objectives. WP0 involves on-going design, scoping and training activities, and involves all members of the project team. Brief updates on WPs 1-5 are provided below.

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

Much progress has been made on this work package over the last few months. A list of transcription conventions, together with a list of anonymisation conventions to be used when building the CorCenCC corpus have been drafted, circulated to experts in the field for feedback, and finalised. Furthermore, our work on the sampling framework, in respect of responding to feedback from Welsh language and corpus experts, is nearing completion.

Aim: Develop the part-of-speech tag-set/tagger (lead: Dawn Knight)

In WP2, we have been setting up an annotation task to provide us with hand-checked data to evaluate our Welsh language processing tools against later.

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

WP3 has focused on developing a prototype of the Welsh semantic tagger, including the Welsh semantic lexicon construction and a software framework. The semantic lexicon currently contains 16,416 Welsh lemmas classified under USAS semantic categories which cover approximately 174,780 inflectional forms of Welsh words. An initial evaluation shows that the current semantic lexicon covers about 72.42% of the words in running Welsh texts.

Aim: Scope/construct the online pedagogic toolkit (leads: Enlli Thomas)

A questionnaire was piloted with a small number of practitioners in the field of Welsh language teaching and, based on their feedback, a new questionnaire was developed for a larger audience. This was distributed to Welsh teachers and tutors at two conferences in July, and an online version will also be produced. This survey – along with focus groups later this year – will help us to identify the priorities for CorCenCC's pedagogical toolkit.

Aim: Construct infrastructure to host CorCenCC and build the corpus (lead: Irena Spasić) In WP5, we have been working on designing the background infrastructure for our corpus, setting up our main project websites, and getting our crowdsourcing application ready for release.

Meet the team

Every month we will be featuring a different member of the extended CorCenCC team in our newsletter. This will give everyone a chance to tell you a little about their background; what they want to see from CorCenCC and how they think it might contribute to their own work, or more broadly, to the work of others in Wales. This month we spotlight Mark Stonelake, who is a Co-Investigator (CI) at Swansea University.

Profile: Mark Stonelake

There's a real need for this. Why hasn't it been done before? Those were my first thoughts on hearing that a new comprehensive Welsh Corpus was on the cards. I was surprised and delighted to be asked to play a small part in the project.

I was born and brought up in Aberdare in the South Wales Valleys. I came to Swansea in 1974 at the tender age of 17, to attend the Art College. Following an unsuccessful attempt at becoming an artist, I got a series of 'proper' jobs as: a farm worker, groundsman, gardener and tree surgeon, before returning to education as a mature student to do a Welsh degree at Swansea University in 1986. I began as a part-time Welsh for Adults tutor at the university at about the same time. After a brief period as a researcher, I became a full time tutor/organiser in 1993, specialising in developing, organising and teaching highly intensive Welsh courses. Between 1996 and 2014, I was also a tutor or the lead tutor on annual courses organised in the USA by Cymdeithas Madog. In 2006 I was appointed as the Curriculum and Resources Officer in the new South-West Wales Welsh for Adults Centre. In a re-structuring last year, I was appointed to the Programme and Curriculum Manager's job. Following further re-structuring, due to big changes in the field this year, I will start a new post as Professional Development and Quality Manager this month in the newly formed Learn Welsh Swansea Bay Region in Academi Hywel Teifi. The curriculum will still be part of my brief, as well as organising the programme of Welsh courses, quality management and training.

Over the years, I have been involved in developing a wide range of resources for Welsh tutors and students, including online and blended courses at all levels. I have been involved in projects with various institutions such as: Y Lolfa, the BBC and S4C, producing resources from course books and graded readers to DVDs and CDs. Most recently, I have been developing material for the S4C app which is used with the programme for Welsh learners *Dal Ati*. I hope my experience in resources development will be useful to the team creating the pedagogical toolkit.

The next generation

The new corpus has the potential to change not only the content of Welsh courses, but also the way the language is taught. Frequency lists will enable the prioritisation of the most used and useful words and phrases. Courses targeting particular age groups, social groups, professions or geographical areas can be developed, based on accurate research, and learners will be able to see how words are used in the correct context. Courses based on the language people actually use, rather than the language we think is used, will be a massive step forward in the teaching of Welsh. It will be a fantastic resource for course designers, researchers, tutors and learners. Not only that, but in future, when I hear: "We don't say it like that" from some know-it-all. I'll be able to say, "Yes you do" and prove it. Now that's what I call progress!

Mark Stonelake, e.m.stonelake@swansea.ac.uk

CorCenCC online

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 holding website at: http://sites.cardiff.ac.uk/corcencc/

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; **Consultants** - Kevin Donnelly, Kevin Scannell, Laurence Anthony, Tom Cobb, Michael McCarthy and Margaret Deuchar; **Project Advisory Group** – Colin Williams, Karen Corrigan, Llion Jones, Maggie Tallerman, Mair Parry-Jones and 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