CorCenCC Newsletter

Issue 12: July 2017

Corpws Cenedlaethol Cymraeg Cyfoes
National Corpus of Contemporary Welsh

Greetings from the PI

Welcome to the twelfth CorCenCC newsletter! The summer is upon us again, so I imagine that many of you will be packing your suitcases and turning on your out of office, ready for a well-deserved break! While the buzz of students and exams has died down at our academic institutions, the buzz on the CorCenCC project continues to be heard throughout the offices of those working in the team. Data collection is on-going and there is a range of public events that members of the team will be attending throughout the summer months, to spread the word on the project and to ask members of the public to 'Give us your Welsh'. If you see us, please come and say hello!

This issue of the newsletter includes reports on recent and future changes in personnel on the project, an update on recent conference activities, and a chance to meet yet another member of the extended team – this time CorCenCC consultant Professor Kevin Scannell from Saint Louis University.

Happy reading and have a lovely summer, Dr Dawn Knight (Cardiff University)

News

Job opportunity

Research Assistant Dr Gareth Watkins will be leaving the project at the end of September (we will all be very sad to see Gareth go) and we will, therefore, be recruiting a new RA to be based at Cardiff University. This is a 23-month post, beginning on 1st October 2017. Details of the role can be found here: http://www.jobs.ac.uk/job/BDB753/research-assistant/ The deadline for this is 11th August 2017. Please circulate the details of this position to anyone who you think might be interested in applying!

New appointment

We would like to say a big farewell to Lorena Varona, CorCenCC's admin assitant, who has recently left the CorCenCC team and moved back home to Spain. All the best for the future Lorena! Replacing Lorena will be Lowri Williams. We asked Lowri to introduce herself to the CorCenCC community…

I'm Lowri and I'm originally from the small town of Pwllheli, North Wales. I was raised in a Welsh speaking home and received my education through the medium of Welsh as a first language. I made the move to the big capital to study Information Systems at the School of Computer Science and Informatics, Cardiff University. Following my BSc, I progressed to read a PhD at the School and was

supervised by Prof. Irena Spasic, who is also a member of the CorCenCC team. My research revolves around the role of idioms in sentiment analysis. I have recently submitted my thesis and await my viva.

As a young native Welsh researcher, the work at CorCenCC is important to me. Not only will it support the Welsh language in the future, but it also has the potential to form a strong basis for research into the development and integration of the language, particularly from a computer science and computational linguistics perspective.

Welcome Lowri - we look forward to working with you on the CorCenCC project!

Would you be interested in working as a CorCenCC transcriber?

As you know, we have been busy recording Welsh being spoken up and down the country. Work has begun on transcribing the recordings, but we are now looking for more transcribers – would you be interested? The work is flexible (you can work whenever suits you, and do as many/few hours as you wish) so it is easy to fit in around other activities, and the recordings are interesting and varied – one day you might be transcribing a lecture or sermon, and the next day a lively conversation down the pub! If you'd be interested in joining our team of transcribers, please email trawsgrifio@corcencc.org for more information.

The CorCenCC launch video – now available online!

We are also pleased to announce that a full version of the CorCenCC launch from March 2017 is now available to view online. To watch the event please visit here: http://www.swansea.ac.uk/cy/riah/prosiectau-ymchwil/corcencc/

Recent events

Tafwyl, Gwyl Fach y Fro, Parti Ponty, and the National Centre for Learning Welsh Conference

As spring turned into summer, as the dark evenings retreated and everyone reached for t-shirts, shorts and sunscreen, we also entered the Welsh language festival season. CorCenCC was fortunate enough to experience some of these festivals first hand, and through doing so met many Welsh speakers. These people were very eager to hear about the project, and were more than willing to contribute their voice to our corpus. From the dazzling sunshine of Tafwyl and Gwyl Fach y Fro, to the refreshing rain of Parti Ponty, people's enthusiasm for the language and for the project was evident.

This time of year also sees the beginning of the conference season. CorCenCC had the opportunity to attend and record at the National Centre for Learning Welsh Conference in July. One of the CorCenCC project's main aims is to create pedagogical tools, that is resources designed to help learners and teachers as they take on the task (pleasure?) of learning and teaching Welsh. This is a wonderful example of potential corpus users contributing to the creation of the resource in the first place.

Massive thanks go out to all those who have contributed to this important work while attending these events. We are all looking forward to meeting and recording yet more Welsh speakers in the coming months.

Gareth Watkins

Corpus Linguistics 2017

On Monday 24th July various members of the CorCenCC team jumped on the train – due North East – to attend the Corpus Linguistics (CL2017) conference at the University of Birmingham. The conference, which comprised one workshop day and 4 main conference days, provided us with the opportunity for disseminating some of the ideas, developments and vision for CorCenCC to other academics and specialists within the field of Corpus Linguistics.

The first paper, presented during the workshop, was an overview paper outlining the main aims and objectives of CorCenCC, whilst discussing some of the key challenges faced in the research. This paper was delivered by Dawn Knight and Paul Rayson, but included a total of 19 co-authors (the main project team members) – certainly a record for CL2017!

Second to the floor were Mair Rees and Gareth Watkins (pictured left), presenting an update on work relating to WP1 and outlining the blueprints for the CorCenCC sampling frame. Thursday morning saw Steve Neale deliver his paper on the CorCenCC crowdsourcing app (WP5), followed by a presentation which asked 'How will you Make Sure the Material is Suitable for Children?: User-Informed Design of Welsh Corpus-Based Learning/Teaching Tools', delivered by Jennifer Needs (WP4) that afternoon. Last but not least, Paul Rayson and Scott Piao presented their work on the CorCenCC semantic tagger (WP3 – pictured above) on the last day of the conference. 5 papers in all! Impressive (but exhausting) stuff!

It was a very thought-provoking, engaging and enriching conference for all involved. We were even treated to some live music by the 'in-house' band 'Corpus Interruptus' (see left)… and even the Midlands wind and rain didn't put the delegates off. At the conference closing we made a very exciting announcement: that the next Corpus Linguistics conference (to be held in 2019) will be hosted by the Centre for Language and Communication

Research at Cardiff University! As this conference coincides with the final year of the CorCenCC project, we intend to incorporate many CorCenCC-related events and activities into the conference, including special Welsh language events – so please look out for details of these. We look forward to seeing many of you at CL2019!

Dewin a Doti and Corpws

Gareth Watkins and Mair Rees went over to St Fagan's on Monday 12th June to join in the fun at the Dewin and Doti Festival which was organised by Mudiad Meithrin (Welsh early years specialists). Dewin is the Mudiad's character and Doti, his dog, is his faithful friend. They are used to introduce and promote the Welsh language to pre-school children. This is a travelling festival, held annually in centres throughout Wales. The festival is an opportunity to bring children and staff of the local Welsh-medium pre-school playgroups and parent and toddler groups together to have fun whilst using their Welsh. As part of the festival, the ever youthful Martyn Geraint presented a show which included the children's favourite songs. Martyn was kind enough to allow CorCenCC to record his session for inclusion in the corpus.

However, one other member of the CorCenCC team also joined in the fun: Corpws the cat! This is a character who helps CorCenCC's staff explain the project to young children, and assists in asking them for their informed consent (along with the consent of their parents) to contribute their spoken Welsh to the project. Here's a picture of Corpws meeting her friend Doti!

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 Professor Kevin Scannell, who is a consultant on the CorCenCC project, based at Saint Louis University.

Profile: Kevin Scannell

First I'd like to say how pleased I am to be able to contribute to the CorCenCC project as an advisor, and I've very much enjoyed reading the other "Meet the Team" pieces!

I'm a Professor of Computer Science at Saint Louis University in the United States, where I've taught since 1998. I am one of a small (but growing!) number of Americans who speak Irish (I began learning the language properly in my 20s), and these days most of the work I do involves developing Irish language technology of one kind or another. I've worked on spelling and grammar checkers, dictionaries, thesauri, machine translation engines, corpora of various kinds, and have done quite a bit analysing the language in social media. My academic background is a bit unusual for this line of work; I was trained as a mathematician and moved

through the academic ranks by publishing research on very abstract (== not-at-all-useful) mathematics and theoretical physics. Fortunately, I came to my senses around 2005 and began devoting all of my research time to projects related to Irish and other indigenous/minority languages. Fundamentally, I believe that everyone should be able to use the computer 100% in their native language; corpus projects like CorCenCC and the research and technologies that stem from them are what will make this a reality for Welsh, Irish, and many other languages.

Almost all of my projects are driven by corpora and statistical methods, going all the way back to about 1999 when I started working on the "GaelSpell" spell checker for Irish. To put it bluntly, I had no idea what I was doing back then, other than to recognize that if I wanted to produce a large word list for the language, then I should start by assembling a large corpus of texts and do some cleaning. Fortunately, there were already quite a few websites in Irish at the time, as well as a vibrant online community centered around email listservs like Gaelic-L and Gaeilge-A (these were the "social networks" of the day). I wrote some quick shell scripts to scrape the text from all of these sites, figured out how to do simple statistical language identification, and by 2004 had built up a corpus of about 20 million words of Irish.

As it happens, Welsh was the first language other than Irish for which I built a web-crawled corpus, collaborating with Dewi Evans from UCD and Andrew Hawke from the Geiriadur Prifysgol Cymru on a big web-crawled corpus of Welsh starting in February 2004. Porting things over to Welsh involved a good bit of refactoring of the crawler to eliminate any "Irish" assumptions and make the whole thing more language independent. So began a slightly obsessive ten-year quest to crawl as many different languages as possible; I've managed to deploy the web crawler and build corpora for more than 2100 different languages as part of the so-called "Crúbadán" project (http://crubadan.org/).

Again, I'm very happy to be contributing to this exciting project, and I'm looking forward to getting to know the rest of the team!

Kevin Scannell

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 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, 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