CorCenCC Newsletter

Issue 15, January 2018

CorCenCC

Corpws Cenedlaethol Cymraeg Cyfoes

National Corpus of Contemporary Welsh

EUROPEAN LANGUAGE RESOURCES ASSOCIATION

ELRA

Contents

Greetings from the PI

Happy New Year to all our readers! I hope you all had a lovely Christmas break (long ago now, though it might seem!) and have had a productive start to 2018. Things have kicked off well here at CorCenCC HQ, and across partner institutions, and plans for a busy year ahead are afoot. Over the next 20 months of the project the fruits of our labour will hopefully become a lot more visible to all, with early releases of the corpus and its associated tools scheduled. A range of public roadshows and events are planned to help us profile and demo CorCenCC – so do keep checking future editions of the newsletter, the Facebook page, Twitter feed and website for more details, and do come and join in! We will also be concentrating on disseminating work at a range of conferences and meetings and hope to have a presence at the Eisteddfod and Tafwyl again, so maybe we will catch you at one of those events too!

The present, slightly compact, edition of the newsletter brings you up-to-date with recent news and updates from the project. This includes details about future travels to France and Japan; some updates on where we are with transcription and web scraping (for the collection of e-language data) and details about our upcoming Whole Project Meeting. As ever, you will get a chance to meet another member of the team (Paul Rayson) and, sadly, we also have to say goodbye to one of our co-investigators on the project, Mark Stonelake, as he steps into retirement.

Happy reading! Dr Dawn Knight

CorCenCC en France

Project lead, Dawn Knight, has been invited to deliver a plenary paper at the Association Francaise de Linguistique Cognitive (The French Cognitive Linguistics Association – AFLiCo) workshop on Corpora and Representativeness. This will be held in Nanterre on 3-4th May 2018. The workshop seeks to explore what is meant by representativeness in corpus design; whether it is something that can ever truly be achieved; how we can address bias in corpora and whether representativeness is the same as balance. Dawn's paper will be entitled 'Representativeness in CorCenCC: corpus design in minoritised languages'

+ News and events

The second CorCenCC Whole Project Team Meeting will take place at Cardiff University on 26th February 2018. All members of the 30-strong CorCenCC project team are invited to this bilingual meeting, which will give attendees an opportunity to track developments across all work packages on the project and to contribute to future planning. We will be discussing a range of topics including how we might generate and capture 'impact' from CorCenCC; plans for future-proofing the corpus; what the best routes to engaging future users are, and what future extensions and satellite projects of CorCenCC are possible. We look forward to seeing all the team in Cardiff!

Whole Project Team Meeting @ Cardiff University

LREC 2018

Members of the CorCenCC team have had the following papers accepted for presentation at the 11th biennial Language Resources and Evaluation Conference (LREC), which will take place from 7-12th May in Miyazaki, Japan:

- Piao, S., Rayson, P., Knight, D. and Watkins, G. (2018). Towards a Welsh Semantic Annotation System. Proceedings of the *LREC (Language Resources Evaluation) 2018 Conference,* May 2018, Miyazaki, Japan.
- Neale, S., Donnelly, K., Watkins, G. and Knight, D. (2018). Leveraging Lexical Resources and Constraint Grammar for Rule-Based Part-of-Speech Tagging in Welsh. Proceedings of the *LREC (Language Resources Evaluation) 2018 Conference,* May 2018, Miyazaki, Japan.

LREC is a highly competitive and prestigious conference that focuses on ground-breaking computational and corpus methods and resource development. It will feature workshops, posters and papers on a variety of themes from sentiment analysis and emotion recognition and corpora for language analysis, to computer-aided language learning. Abstracts are double blind peer-reviewed, and have a low rate of acceptance. As this is a computational linguistics conference, each of these submissions comprises an extended abstract and will lead to publication in the proceedings of the conference. Attending the conference will enable us to present some key findings derived from our work on WP3 and WP2 of CorCenCC to the wider computational/corpus community, and to profile the aims and objectives of the project more widely. We also hope to catch up with project consultant Laurence Anthony while we are in Japan. We will bring you a report of how we get on at this conference, later in the summer.

Mark Stonelake retires

We are very sad to announce that Mark Stonelake, Co-Investigator on the CorCenCC project, will be leaving the team as he retires from Swansea University. As a materials writer and experienced language teacher, Mark has been a very valuable member of the project team; we will miss the insights and enthusiasm he brings to our team meetings. We want to say a big thank you for all his hard work, and wish him all the best for his retirement. We know Mark's enthusiasm for the aims and ethos of the CorCenCC work will continue, and we hope he will stay involved in the project as time goes on - don't be a stranger Mark! We asked Mark to provide a few words on his retirement – this is what he said: 'Is it that time already? Only a few years ago, started a degree course at Swansea University. Well, that's how it feels, anyway. But that was back in 1986! I'd just given up my job as a tree surgeon/gardener to do a Welsh BA. In those days, students used to get a grant, not a loan and I, as a mature student, got an enhanced grant. Ah, the good old days! Had there been no grants back then, and us trying to raise a young family, I don't think I'd have taken the risk. Even so, money was tight, so when I was asked if I fancied teaching a night class, I jumped at the chance. People didn't seem to worry too much about qualifications back in the day. Since then, I've worked, with a brief period out as a part-time tutor and house husband (is that the correct term these days?), in various full-time posts for Welsh for Adults.

I've thoroughly enjoyed all aspects of the work, from teaching and materials development to training and managing staff. Leading many courses in Wales and America and visits to Finland, Catalonia and the Basque country has been fantastic. I've had the opportunity to meet people from many different backgrounds and nationalities on the various courses, and as part of the CorCenCC team. I would never have been able to do this had I not taken the risk of going to University and changing career back in the 80s.

I hope I've made a valuable contribution to the field of teaching Welsh to Adults over the years, and I hope I've also made a small contribution to CorCenCC. However, the time has come to put away the white board and spreadsheets. I'll be retiring from my job as Professional Development and Quality Manager at the end of February. I'll remember our achievements and hard work. I'll remember my students and colleagues fondly. Well, most of them, anyway. But the thing I'll remember most is the laughter and 'hwyl'. Most of the time, I've had such fun, it didn't seem like work at all.

I hope to teach the odd Welsh class (not the normal ones) as an associate tutor and still be involved in some way with the development of the corpus. So, you probably haven't seen the last of me yet.

So, Mark, with your vast knowledge and great wisdom, gleaned from years of back-breaking work on the front line of education, do you have any advice for us? Well, since you ask, I think that we should all play close attention to the wise words of 'Mr New Year' himself - Guy Lombardo:

Enjoy yourself, it's later than you think
Enjoy yourself, while you're still in the pink
The years go by, as quickly as you wink
Enjoy yourself, enjoy yourself, it's later than you think

I've thoroughly enjoyed working with you all and wish you every success and happiness for the future.

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

Data collection and participant recruitment

We're hoping to really hit the ground running in 2018 and spread the word about what's been happening here on CorCenCC. Social media will play an important role, and we've been thinking hard about how best to use our existing Facebook and Twitter channels to do this. So, what's been happening… and what's coming next?

Sharing is caring - The great thing about contributing to the project is that it can be done from home on your mobile phone. Gone are the days where you needed fancy technology to record (although we do still use these and they can be a great help!). Now all you need is an iPhone and somewhere (and someone) to record. So, we encourage you, via Facebook and Twitter, to download the iPhone app and get involved. Look out for the hashtags #WeWantYourWelsh #RhowchEichCymraegINi and please help us by sharing our posts. The more the merrier!

A growing team - We're always on the lookout for new transcribers, and social media has been a great recruitment tool. We put out a call for transcribers on Facebook before Christmas and again on Twitter in the last couple of weeks and the uptake has been great. To date we've had 60 people show interest in working for us, with new requests every day. We're really excited about what we can achieve together this year!

Coming soon to a coffee shop/hairdressers/school near you… - our Facebook and Twitter pages (and this newsletter) are great way to find out what the CorCenCC team are up to.

We'll let you know if we're coming to a particular area or event that might interest you. Is there someone we simply have to chat to? Is there an event you think would be great for the corpus? Let us know!

So, keep an eye out for lots going on this year, especially on our social media accounts. The CorCenCC project is all about collaboration and innovation, we'd love for you to get involved. Follow the CorCenCC project on Facebook or on Twitter @corcencc. **Lowri and Lowri**

+ General update: E-language collection

Web-scraping in progress!

By Lowri Williams

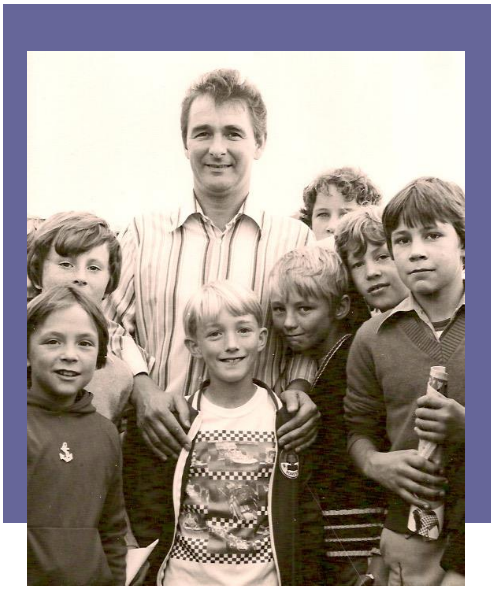
Since our last update, one of the main focuses for WP1 has been on collecting data from websites and other types of electronic text. Our goal is to collect 2 million words used in such contexts. The first task was to identify websites or blogs which presented their text through the Welsh medium. We have been contacting website owners or bloggers to ask for their consent to include their material in our corpus. We currently have a total of 152 websites and blogs who have kindly agreed to contribute their contexts to CorCenCC. A large number of sites means that there is a large amount of data to be collected. Given that my background is in computer science, I was able to utilise my skills to implement a web scraper to automatically extract texts from such resources. The web scraper is written in Python, a high-level programming language. The first requirement is to identify which pages of a given website are to be used. Using a sequence of Python libraries, we are able to enter a website URL and return all the remaining URLs for the given site. These URLs are then inputted to the scraper which returns and cleans the HTML, the language for creating web pages. We are able to retrieve text from a website in less than 5 minutes. The current total number of words that have been scraped is 1.5 million. The next task is to anonymise the data to be incorporated into the final corpus.

+ Meet the team:

Dr Paul Rayson, Co-Investigator, Lancaster University

So, I should start with where I'm from. I was born in the sixties (before Neil Armstrong walked on the moon!) and lived near Nottingham and went to school at Toot Hill Comprehensive in Bingham, great name for a school! But now I've spent longer living in Lancaster after arriving as an undergrad student in 1987, then staying on as a research assistant and PhD student, then a teaching fellow, lecturer, senior lecturer and now reader in natural language processing. Before I was 10, my granddad got me started supporting Nottingham Forest who were and obviously still are the best team in the world. In those days, we'd won two European cups back to back and we even followed them to Wembley for a League Cup final. Given the team I support, it's a lot of fun to be working with other CorCenCC project team members, one of which is a Derby County fan (boo hiss!!), and another who has played for the Forest Ladies team (hooray!). Can you guess who? If not, have a look back at the 'Meet the Team' sections in previous newsletters!!

I first became interested in computers at school in the heady days of ZX Spectrums and BBC Micros. It was a mix of playing games and typing programs in from magazines or loading them in from cassette tapes (kids – ask your parents what a tape is!). We basically ran the computer lab at school for the teachers and I received a BBC Micro as a joint birthday and Xmas present one year. Back in those days, BBC Micros had 64K of memory. Who could possibly need more than that?! Now, I have emails larger than 64K. I ended up at Lancaster as they were one of only five UK universities at the time running a joint Computer Science and Maths degree, plus they offered me a £70 scholarship. £70 was a lot of money in 1987!! My route into natural language processing started when I chose a third year undergraduate project building a CYK parser supervised by Roger Garside, and then he offered me a job (no interviews needed in those days!) before my finals in 1990.



Fast forward to Tuesday 24th September 2013 when Dawn invited me to get on a train (well, several trains actually) to Cardiff for a CorCenCC proposal launch meeting at the Senedd building. The rest as they say, is history. I'd been working on NLP and in particular semantic taggers, originally for English in 1990, but for many other languages since then, so it's been really interesting creating a semantic tagger for Welsh in the CorCenCC project along with Scott Piao. Previously, my only Welsh vocabulary was on the road signs I'd seen on the way to/from holidays in Wales. Obviously, the other major attraction to working on the project has been the chance to get proper Welsh cakes on a regular basis!! Oh, and plus the Doctor Who Experience was based in Cardiff, so I really couldn't say no, could I!

+ 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 Lowri Williams; 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, 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